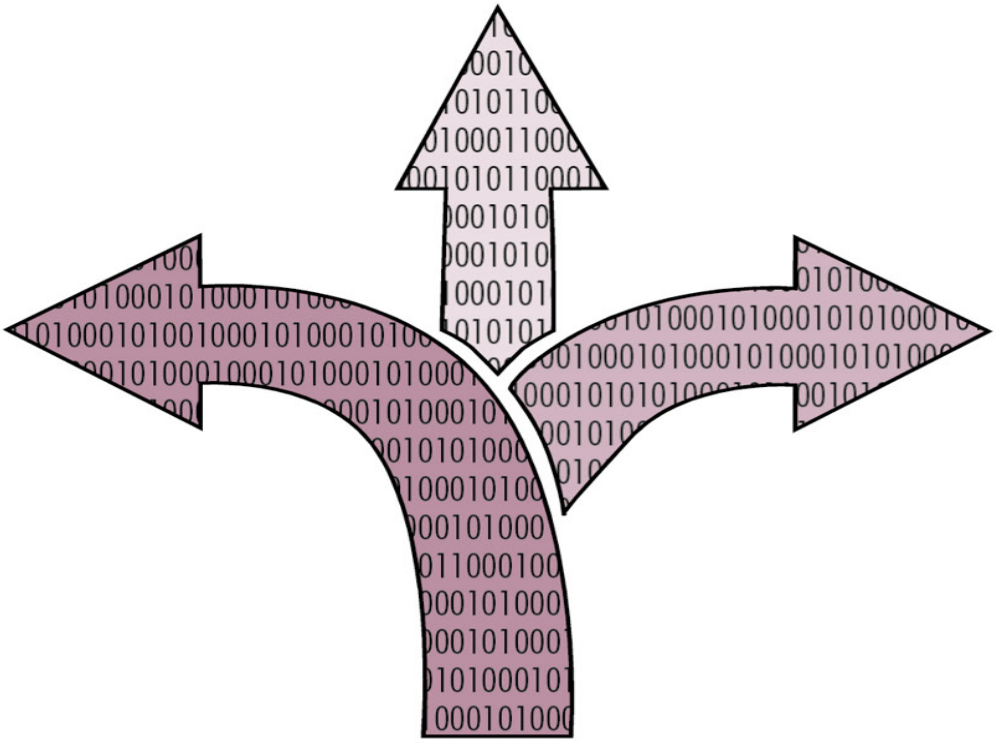
ÉTICA DE LA INTELIGENCIA ARTIFICIAL

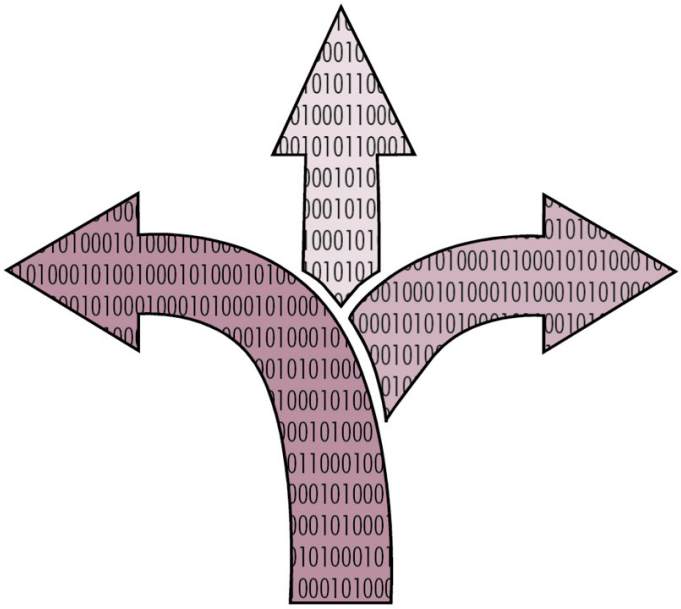
Mark Coeckelbergh



CÁTEDRA

ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Mark Coeckelbergh



CÁTEDRA

Mark Coeckelbergh

Ética de la inteligencia artificial

Traducción de Lucas Álvarez Canga

CÁTEDRA

Índice

Para Arno

Agradecimientos

Este libro no solo se basa en mi propio trabajo en esta materia, sino que también refleja el conocimiento y la experiencia de todo el campo de la ética en la inteligencia artificial. Sería imposible hacer una lista con toda la gente con la que he discutido y aprendido a lo largo de los años, pero en las relevantes y cada vez más numerosas comunidades que conozco se incluyen investigadores en inteligencia artificial [IA] como Joanna Bryson y Luc Steels, compañeros filósofos de la tecnología como Shannon Vallor y Luciano Floridi, académicos que trabajan en innovación responsable en Países Bajos y Reino Unido como Bernd Stahl en la Universidad De Montfort, gente que conocí en Viena como Robert Trappl, Sarah Spiekermann y Wolfgang (Bill) Price, y mis compañeros de los órganos consultivos orientados a la política pertenecientes al Grupo de Expertos de Alto Nivel en IA (Comisión Europea) y al Consejo austriaco de robótica e inteligencia artificial, por ejemplo Raja Chatila, Virginia Dignum, Jeroen van den Hoven, Sabine Köszegi y Matthias Scheutz, por nombrar unos pocos. También me gustaría dar afectuosamente las gracias a Zachary Storms por ayudarme con la corrección y la preparación del libro, y a Lena Starkl e Isabel Walter por ayudarme con la búsqueda de bibliografía.

Capítulo 1

Espejito, Espejito

La exageración de la IA y sus miedos: espejito, espejito, ¿quién es el más listo del reino?

Cuando se anunciaron los resultados, los ojos de Lee Sedol se llenaron de lágrimas. AlphaGo, un programa de inteligencia artificial (IA) desarrollado por DeepMind de Google consiguió la victoria por 4 a 1 en el juego de Go. Es marzo de 2016. Dos décadas antes, el gran maestro de ajedrez Garry Kasparov perdió contra la máquina Deep Blue, y ahora un programa de ordenador ganó contra el dieciocho veces campeón del mundo Lee Sedol en un complejo juego al que se creía que solo los humanos podían jugar, utilizando su intuición y su pensamiento estratégico. El ordenador ganó siguiendo no solo las reglas dadas por los programadores, sino empleando también un sistema de aprendizaje automático basado en millones de partidas anteriores de Go y jugando contra sí mismo. En este caso, los programadores prepararon las bases de datos y crearon los algoritmos, pero no podían saber cuáles serían los movimientos que elaboraría el programa. La IA aprende por sí misma. Después de varias jugadas inusuales y sorprendentes, Lee tuvo que rendirse (Borowiec 2016).

Se trata de un logro increíble para la IA, pero también suscita inquietudes. Hay una admiración por la belleza de las jugadas, pero también tristeza, incluso miedo. Existe la esperanza de que IAs más inteligentes puedan incluso ayudarnos a revolucionar los servicios sanitarios o a encontrar soluciones para todo tipo de problemas sociales, pero también la preocupación de que las máquinas tomen el control. ¿Podrán las máquinas superarnos en inteligencia y controlarnos? ¿Es la IA una mera herramienta, o poco a poco y de forma segura se está convirtiendo en nuestro amo y señor? Estos mie-

dos nos recuerdan las palabras del ordenador HAL en la película de ciencia ficción de Stanley Kubrick *2001: Una odisea en el Espacio,* quien en respuesta a una orden humana, «abre las puertas del hangar», responde: «lo siento Dave, me temo que no puedo hacer eso». Y si no es miedo, entonces puede que sea un sentimiento de tristeza o decepción. Darwin y Freud destronaron nuestras creencias de que éramos excepcionales, nuestros sentimientos de superioridad y nuestras fantasías de control; hoy día, la inteligencia artificial parece asestar otro golpe a la autoimagen de la humanidad. Si una máquina puede hacer esto, ¿qué queda para nosotros? ¿Qué somos? ¿Somos simplemente máquinas? ¿Somos máquinas *inferiores,* con demasiados defectos? ¿Qué será de nosotros? ¿Nos convertiremos en los esclavos de las máquinas? O, peor, ¿en una mera fuente de energía, como en la película *Matrix?*

El verdadero y generalizado impacto de la IA

Pero los avances de la inteligencia artificial no se limitan a los juegos o al mundo de la ciencia ficción. La IA se da ya en la actualidad, y está generalizada, a menudo integrada de forma invisible en nuestras herramientas cotidianas como parte de complejos sistemas tecnológicos (Boddington 2017). Dado el crecimiento exponencial de la potencia de los ordenadores, la disponibilidad de grandes conjuntos de datos debida a las redes sociales y al uso masivo de miles de millones de smartphones y redes móviles de gran velocidad, la IA, y especialmente el aprendizaje automático, ha logrado avances significativos. Este hecho ha permitido a los algoritmos hacerse cargo de muchas de nuestras actividades, incluyendo la planificación, el habla, el reconocimiento facial y la toma de decisiones. Las aplicaciones de la IA se dan en muchos ámbitos, incluyendo transporte, *marketing,* servicios sanitarios, finanzas y aseguradoras, la seguridad y el ámbito militar, ciencia, educación, trabajo de oficina y asistencia personal (por ejemplo, Google Duplex) , entretenimiento, artes (recuperación de información musical y composición), la agricultura y, por supuesto, la fabricación.

La IA se da ya en la actualidad, y está generalizada, a menudo integrada de forma invisible en nuestras herramientas cotidia-nas.

La IA es creada y utilizada por empresas de tecnología de la información [TI] y de internet. Por ejemplo, Google siempre ha usado la IA para su motor de búsqueda. Facebook usa IA para la publicidad dirigida y el etiquetado de fotos. Microsoft y Apple usan IA para potenciar sus asistentes digitales. Pero la aplicación de la IA abarca mucho más que el sector TI definido en sentido estricto. Por ejemplo, hay muchos planes concretos y experimentos con coches autónomos. Esta tecnología también está basada en la IA.

Los drones usan IA, así como las armas autónomas que pueden matar sin intervención humana. Y la IA ya se ha empleado para tomar decisiones en juzgados. En los Estados Unidos, por ejemplo, el sistema COMPAS se ha utilizado para predecir quién es más probable que vuelva a delinquir. La IA entra también en campos que normalmente consideramos que son más personales o íntimos. Por ejemplo, las máquinas ahora pueden leer nuestras caras: no solo para identificarnos, sino también para interpretar nuestras emociones y recuperar todo tipo de información.

La necesidad de discutir los problemas éticos y sociales

La IA puede tener muchos beneficios. Se puede usar para mejorar los servicios públicos y comerciales. Por ejemplo, el reconocimiento de imágenes es una buena noticia para la medicina: puede ayudar en el diagnóstico de enfermedades como el cáncer o el alzhéimer. Pero las aplicaciones cotidianas de la inteligencia artificial muestran también que las nuevas tecnologías plantean problemas éticos. Daré algunos ejemplos de conflictos éticos relacionados con la IA.

¿Deberían tener los coches autónomos restricciones éticas y, si así fuera, qué tipo de restricciones, y cómo deberían determinarse? Por ejemplo, si un coche autónomo se encuentra en una situación en la que debe escoger entre atropellar a un niño y chocar contra un muro para salvar la vida del niño, pero potencialmente matar a su pasajero, ¿qué debería escoger? Y ¿deberían estar siquiera permitidas las armas letales autónomas? ¿Cuántas decisiones y cuánto de estas decisiones queremos delegar a la IA? Y ¿quién es respon-

sable cuando algo sale mal? En cierto caso, los jueces depositaron más confianza en el algoritmo COMPAS que en los acuerdos a los que llegaron la defensa y el fiscal . ¿Se confía demasiado en la IA? El algoritmo COMPAS es también enormemente controvertido, puesto que la investigación ha demostrado que los falsos positivos del algoritmo (personas que predijo que iban a volver a delinquir pero que no lo hicieron) se daban desproporcionadamente entre gente de piel negra (Fry 2018). La IA puede, así, reforzar prejuicios y discriminaciones injustas. Problemas similares pueden surgir con algoritmos que recomiendan decisiones sobre solicitudes de préstamos y de empleo. O considérese la llamada policía predictiva: los algoritmos se usan para predecir dónde es probable que ocurran los delitos (por ejemplo, en qué área de una ciudad) y quién puede cometerlos, pero el resultado puede ser que un grupo específico socioeconómico o racial esté señalado desproporcionadamente por la vigilancia policial. La policía predictiva ya se ha usado en los Estados Unidos y, según lo que muestra un informe reciente de AlgorithmWatch (2019), también en Europa . Y la IA basada en la tecnología de reconocimiento facial se utiliza a menudo para la vigilancia y puede violar la privacidad de las personas. También puede predecir de manera más o menos aproximada las preferencias sexuales. No se necesita información del número de teléfono ni datos biométricos. La máquina hace su trabajo a distancia. Con cámaras en la calle y en otros espacios públicos, podemos ser identificados y «leídos», incluso por lo que respecta a nuestro estado de ánimo. Mediante el análisis de nuestros datos, se puede predecir nuestra salud mental y corporal (sin que lo sepamos). Los empresarios pueden usar la tecnología para monitorizar nuestro rendimiento. Y los algoritmos que están activos en las redes sociales pueden propagar discursos de odio o información falsa. Por ejemplo, los *bots* políticos pueden aparentar ser personas reales y publicar contenido político. Un caso conocido es el del *chatbot* de Microsoft en 2016, llamado Tay y diseñado para tener conversaciones lúdicas en Twitter, pero que, cuando se volvió más inteligente, empezó a *tuitear* contenidos racistas. Algunos algoritmos de IA pueden incluso crear vídeos de discursos falsos, como el vídeo que se compuso para inducir falsamente al público a pensar que parecía un discurso dado por Barack Obama .

Las intenciones normalmente son buenas. Pero estos problemas éticos son normalmente consecuencias no deseadas de la tecnología: la mayoría de estos efectos, como los prejuicios o los discursos de odio, no eran algo que pretendieran hacer los que desarrollan o son usuarios de la tecnología. Además, una cuestión crítica que debe siempre preguntarse es: ¿mejora para quién? ¿El gobierno o los ciudadanos? ¿La policía o aquellos que la policía tiene en su punto de mira? ¿El vendedor o el cliente? ¿Los jueces o los acusados? Las cuestiones relacionadas con el poder entran en acción, por ejemplo, cuando la tecnología se forma solo para algunas pocas mega corporaciones (Nemitz 2018). ¿Quién determina el futuro de la IA?

Esta cuestión subraya el significado social y político de la IA. La ética de la IA se ocupa del cambio tecnológico y su impacto en las vidas de los individuos, pero también de las transformaciones que se producen en la sociedad y en la economía. Las cuestiones que tienen que ver con el prejuicio y la discriminación ya indican que la IA resulta relevante socialmente. Pero también está cambiando la economía y, por tanto, quizás la estructura de nuestras sociedades. De acuerdo con Brynjolfsson y McAfee (2014), hemos entrado en una segunda edad de la máquina en la que las máquinas no son únicamente un complemento de los humanos, como en la Revolución Industrial, sino que también los sustituyen. Ya que profesiones y trabajos de todo tipo se verán afectados por la IA, se ha predicho que nuestra sociedad cambiará drásticamente a medida que pasen al mundo real ciertas tecnologías antaño descritas en la ciencia ficción (McAfee y Brynjolfsson 2017). ¿Cuál es el futuro del trabajo? ¿Qué tipo de vidas tendremos cuando la IA asuma puestos de trabajo? ¿Y quién somos «nosotros»? ¿Quién saldrá ganando con esta transformación, y quién perdiendo?

La ética de la IA se ocupa del cambio tecnológico y su impacto en las vidas de los individuos, pero también de las transformaciones en la sociedad y en la economía.

Este libro

Ciertos avances espectaculares han provocado un gran problema de exageración en torno a la IA. Y esta ya se está utilizando en un amplio grupo de áreas de conocimiento y de prácticas humanas. Lo primero ha hecho surgir especulaciones disparatadas sobre el futuro tecnológico e interesantes discusiones filosóficas sobre lo que significa ser humano. Lo segundo ha creado una sensación de urgencia en parte de los estudiosos de la ética y los políticos para asegurar que esta tecnología nos beneficia en lugar de propiciar desafíos insuperables para determinados individuos y sociedades. Estas últimas preocupaciones son más prácticas e inmediatas.

Este libro, escrito por un filósofo académico que tiene también experiencia en asesoramiento para el establecimiento de algunas políticas, lidia con ambos aspectos: trata de la ética en tanto que relacionada con todas estas cuestiones. Busca dar al lector una buena visión general de los problemas éticos que surgen en conexión con la IA entendida de forma amplia, desde las influyentes narrativas sobre el futuro de la IA y las cuestiones filosóficas sobre la naturaleza y el futuro de lo humano, hasta las preocupaciones éticas sobre la responsabilidad, el prejuicio y el modo de lidiar con cuestiones prácticas del mundo real que surgen de la tecnología mediante la aplicación de políticas (preferiblemente antes de que sea demasiado tarde).

¿Qué pasará cuando sea «demasiado tarde»? Algunos escenarios son distópicos y utópicos al mismo tiempo. Comenzaré con algunos sueños y pesadillas sobre el futuro tecnológico, relatos de gran repercusión que, al menos *a primera vista,* parecen relevantes para evaluar los potenciales beneficios y peligros de la inteligencia artificial.

Capítulo 2

Superinteligencia, monstruos y el apocalipsis de la IA

Superinteligencia y transhumanismo

Esta exageración que rodea a la IA ha hecho surgir todo tipo de especulaciones sobre su futuro y, ciertamente, sobre el futuro de lo que es ser humano. Una idea popular, que no solo se repite a menudo en los medios de comunicación y en el discurso público sobre la IA, sino que también es contemplada por gente influyente del mundo tecnológico que desarrolla tecnología de IA, como Elon Musk y Ray Kurzweil, es la idea de la superinteligencia y, más generalmente, la idea de que las máquinas tomarán el control, que nos dominarán, más que al contrario. Para algunos, esto es un sueño; para muchos, una pesadilla. Y para otros las dos cosas a la vez.

La idea de la superinteligencia es que las máquinas superarán la inteligencia humana. Está a menudo vinculada con la idea de una explosión de inteligencia y de una singularidad tecnológica. De acuerdo con Nick Bostrom (2014), nuestra situación sería comparable con la de los gorilas, cuyo destino depende hoy en día completamente de nosotros. Este autor señala al menos dos caminos hacia la superinteligencia y lo que a veces se denomina *explosión de inteligencia.* Uno es que la IA desarrolle una automejora recursiva: una IA podría diseñar y mejorar versiones de sí misma, que, a su vez, diseñarían versiones más inteligentes, y así sucesivamente. Otro camino es la completa emulación cerebral o su «carga» *(upload)*: un cerebro biológico que pudiera ser escaneado, modelado y reproducido en y por software inteligente. Esta simulación de un cerebro biológico estaría conectada entonces con un cuerpo robótico. Estos desarrollos conducirían a una ex-

plosión de inteligencia no humana. Max Tegmark (2017) imagina que un equipo de investigadores podría crear una IA que se volviera todopoderosa y dirigiera el planeta. Y Yuval Harari escribe sobre un mundo que ya no está gobernado por los seres humanos, sino en el que se adora a los datos y se confía en algoritmos para tomar decisiones. Después de que se hayan destruido todas las ilusiones humanistas e instituciones liberales, los humanos soñarían solamente con fundirse con el flujo de datos. La IA sigue su propio camino, «yendo hacia donde ningún humano ha ido antes, y donde ningún humano puede ir» (Harari 2015, 393).

La idea de una explosión de inteligencia está estrechamente relacionada con la de la *singularidad tecnológica:* un momento en la historia humana en el que el progreso tecnológico exponencial traería consigo cambios tan drásticos que dejaríamos de comprender lo que ocurre y en el que «los asuntos humanos, entendidos como se entienden hoy día, llegarían a su fin» (Shanahan 2015, xv). En 1965, el matemático británico Irving John Good especuló con una máquina ultrainteligente que diseña máquinas mejores; en los 90, el autor de ciencia ficción e informático Vernor Vinge afirmaba que esto implicaría el fin de la era humana. El pionero de la ciencia de la computación John von Neumann ya habría sugerido esta idea en los años 50. Ray Kurzweil (2005) abrazó el término «singularidad» y predijo que la IA, junto con los ordenadores, la genética, la nanotecnología y la robótica, llegarían al punto en el que la inteligencia de las máquinas sería más potente que la de la suma de todos los seres humanos y que, en última instancia, la inteligencia humana y la de las máquinas se fusionarían. Los humanos trascenderían así las limitaciones de sus cuerpos biológicos. Además, como afirma el título de su libro: la singularidad está cerca. Piensa que ocurrirá alrededor de 2045.

Esta historia no tiene necesariamente un final feliz: para Bostrom, Tegmark y otros, hay «riesgos existenciales» vinculados a la superinteligencia. El resultado de estos desarrollos puede ser que una IA superinteligente supere y amenace la vida humana. Sea o no consciente tal entidad, en general, e independientemente de su condición y de cómo llegue a existir, la preocupación con la que nos topamos aquí tiene que ver con lo que esta entidad podría hacer (o no hacer). La IA puede que no se preocupe de nuestros fines

humanos. Al no tener un cuerpo biológico, puede que no llegue siquiera a entender nuestro sufrimiento. Bostrom ofrece el experimento mental de una IA a la que se le da la finalidad de maximizar la fabricación de clips, y lo hace convirtiendo a la Tierra y a los humanos que viven en ella en recursos para producir clips. El desafío que tenemos por delante es, pues, asegurar que construimos una IA que de alguna manera no haga surgir este problema de control: que haga lo que queremos que haga y que tome en consideración nuestros derechos. Por ejemplo, ¿deberíamos limitar de alguna manera las capacidades de la IA? ¿Cómo se supone que la contendremos? .

Un conjunto de ideas relacionado con esto es el del *transhumanismo*. Ante la superinteligencia y la decepción por la fragilidad humana y sus «errores», los transhumanistas como Bostrom sostienen que necesitamos mejorar al ser humano: hacerlo más inteligente, menos vulnerable a enfermedades, con una vida más larga, y potencialmente inmortal: llevándonos así a lo que Harari llama el *Homo deus:* humanos que han ascendido a la categoría de dioses. Como ya dijo Francis Bacon en «La refutación a las filosofías»: los humanos son «dioses mortales» (Bacon 1964, 106). ¿Por qué no tratar de alcanzar la inmortalidad? Pero incluso si no pudiéramos alcanzarla, de acuerdo con los transhumanistas, es necesario mejorar la máquina humana. Si no lo hacemos, los humanos nos arriesgamos a quedarnos como «la parte lenta y cada vez más ineficiente» de la IA (Armstrong 2014, 23). La biología humana necesita rediseñarse y, argumentan algunos transhumanistas, ¿por qué no deshacernos de todas las partes biológicas y diseñar seres inteligentes no orgánicos?

A pesar de que la mayoría de filósofos y científicos que tienen en consideración estas ideas se preocupan por distinguir sus opiniones de la ciencia ficción y de la religión, muchos investigadores interpretan sus ideas precisamente en estos términos. Para empezar, no está claro hasta qué punto sus ideas son relevantes para los actuales desarrollos tecnológicos y para la ciencia de la IA, ni si hay una oportunidad real de que alcancemos la superinteligencia en un futuro previsible; ni siquiera si existe una oportunidad. Algunos rechazan categóricamente esta misma posibilidad (véase el siguiente capítulo), y aquellos preparados para aceptar que es posible en principio, como, por ejemplo, la científica Margaret Boden, no piensan que

vaya a ocurrir en la práctica. La idea de la superinteligencia supone que desarrollemos la llamada *inteligencia artificial general* (o *fuerte)*, o una inteligencia que compita o exceda a la humana, y que existen muchos obstáculos que superar antes de conseguirlo. Boden (2016) ha argumentado que la IA es menos prometedora de lo que mucha gente supone. Y un informe de la Casa Blanca de 2016 respalda un consenso entre expertos del sector privado en el sentido de que la IA general no se alcanzará en, al menos, varias décadas. Muchos investigadores de la IA también rechazan las visiones distópicas que fomentan Bostrom y otros, y subrayan el uso positivo de la IA como ayudante o compañero de equipo. Pero la cuestión no es solamente qué ocurrirá realmente en el futuro. Otra preocupación es que esta discusión sobre los futuros efectos de la IA nos distraiga de los riesgos reales y presentes de sistemas ya instalados. Parece haber un riesgo real de que en un futuro cercano los sistemas ya no sean *lo bastante inteligentes* y de que, aun comprendiendo de manera deficiente sus implicaciones éticas y sociales, los usemos ampliamente. El énfasis excesivo en la inteligencia como característica principal de la humanidad y nuestro fin último es también cuestionable (Boddington 2017).

Sin embargo, ideas como la superinteligencia continúan influyendo en la discusión pública. También es posible que tengan un impacto sobre el desarrollo de la tecnología. Por ejemplo, Ray Kurzweil no es solo un futurólogo. Desde 2012 ha sido director de ingeniería de Google. Y Elon Musk, CEO de Tesla y SpaceX y una figura pública muy conocida, parece apoyar tanto la idea de la superinteligencia como las de los escenarios de riesgo existencial (¿escenarios catastrofistas?) de Bostrom y Kurzweil. Ha advertido repetidamente sobre los peligros de la inteligencia artificial, considerándola una amenaza existencial y afirmando que no podemos controlar al demonio (Dowd 2017). Piensa que los humanos probablemente se extingan, a menos que la inteligencia humana y la de la máquina se fusionen o consigamos escapar a Marte.

Quizás estas ideas tengan tanta repercusión porque tocan preocupaciones y esperanzas profundas respecto de los humanos y las máquinas, que están presentes en nuestra conciencia colectiva. Tanto si rechazamos estas ideas en concreto como si no, hay en la cultura y en la historia humanas claros

vínculos con narrativas de ficción que intentan otorgar sentido a la relación entre humanos y máquinas. Merece la pena hacer explícitas estas narrativas para contextualizar y comprender mejor algunas ideas. De forma más general, es importante incorporar la investigación de narrativas a la ética de la inteligencia artificial: por ejemplo, para entender por qué ciertos relatos prevalecen, quién los crea y quién se beneficia de ellos (Royal Society 2018). Esto también puede ayudarnos a construir nuevas narrativas para el futuro de la IA.

El nuevo monstruo de Frankenstein

Una forma de dejar atrás las exageraciones en torno a la IA es detenerse en algunas narrativas importantes de la historia de la cultura humana que conforman la discusión pública actual. No es la primera vez que la gente se hace preguntas sobre el futuro de la humanidad y el futuro de la tecnología. Y, por exóticas que parezcan algunas ideas sobre la IA, podemos explorar conexiones con ideas y narrativas bastante familiares que están presentes en nuestra conciencia colectiva o, de forma más precisa, en la conciencia colectiva de Occidente.

Primero, hay una larga historia del pensamiento sobre seres humanos y máquinas o criaturas artificiales, tanto en la cultura occidental como en las no occidentales. La idea de crear seres vivos a partir de materia inerte se puede encontrar en las historias de la creación de las tradiciones sumeria, china, judía, cristiana y musulmana. Los antiguos griegos ya tenían la idea de crear humanos artificiales, en particular mujeres artificiales. Por ejemplo, en *La Ilíada,* se dice que Hefesto es asistido por sirvientes hechos de oro que parecen mujeres. En el famoso mito de Pigmalión, un escultor se enamora de la estatua de marfil de una mujer que él mismo ha hecho. Desea que cobre vida, y la diosa Afrodita le concede su deseo: sus labios se vuelven calientes y su cuerpo suave. Es fácil ver aquí la conexión con los robots sexuales de la actualidad.

Estas narrativas no provienen solo de los mitos: en su libro *Automata,* el matemático griego e ingeniero Herón de Alejandría (ca. 10-ca. 70 d.C.) publicó descripciones de máquinas que hacían que la gente creyera que había

visto manifestaciones divinas en los templos; en 1901, se encontró un artefacto en el mar, el mecanismo de Anticitera, que se ha identificado como un ordenador analógico de la Grecia antigua basado en un complejo mecanismo de relojería. Pero las historias ficticias en las que las máquinas parecen ser humanas nos fascinan especialmente. Considérese, por ejemplo, la leyenda del Golem: un monstruo hecho de arcilla creado en el siglo XVI por un rabino que se aparta de la norma. Encontramos en ella una versión primitiva del problema del control. El mito de Prometeo también se interpreta a menudo de esta manera: roba el fuego de los dioses y se lo da a los humanos, pero es castigado por ello. Su tormento eterno consiste en estar atado a una roca a la espera de un águila que todos los días le devora el hígado. La antigua lección consistía en alertar de la soberbia: tales poderes no están destinados a los mortales.

Sin embargo, en *Frankenstein* de Mary Shelley (que lleva el relevador subtítulo de *El moderno Prometeo)* la creación de vida inteligente a partir de materia inerte se convierte en un proyecto científico moderno. El científico Victor Frankenstein crea un ser de aspecto humano a partir de partes de cadáveres, pero pierde el control de su criatura. Mientras que el rabino puede, al final, controlar al Golem, en este caso no es así. *Frankenstein* puede considerarse como una novela romántica que nos alerta de los peligros de la tecnología moderna, pero está informada por la ciencia de su tiempo. Por ejemplo, el uso de electricidad (una tecnología muy joven por entonces) desempeña un papel importante: se usa para animar el cadáver. También hace referencias al magnetismo y a la anatomía. Pensadores y escritores de su tiempo debatían sobre la naturaleza y el origen de la vida. ¿Qué es la fuerza vital? Mary Shelley estaba influenciada por la ciencia de su tiempo . La historia muestra que los románticos del siglo XIX estaban a menudo fascinados por la ciencia, tanto como esperaban que la poesía y la literatura nos liberase de los oscuros entresijos de la modernidad (Coeckelbergh 2017). La novela no debería considerarse necesariamente contra la ciencia y la tecnología: el mensaje principal parece ser que los científicos necesitan asumir la responsabilidad de sus creaciones. El monstruo huye, pero lo hace porque su creador lo rechaza. Es importante tener esta lección en mente para la ética de la IA. Sin embargo, la novela subraya claramente el peligro de la tec-

nología de la que se pierde el control, y en especial de los humanoides artificiales enloquecidos. Este miedo reaparece entre las preocupaciones contemporáneas acerca de una IA descontrolada.

Sin embargo, en *Frankenstein* de Mary Shelley (que lleva el relevador subtítulo de *El moderno Prometeo)* la creación de vida inteligente a partir de materia inerte se convierte en un proyecto científico moderno.

Además, como en *Frankenstein* y en la leyenda del Golem, existe una narrativa que incorpora la competición: la creación artificial compite con la humana. Esta narrativa continúa conformando nuestra ciencia ficción sobre la IA, pero también nuestro pensamiento contemporáneo sobre tecnologías como la IA y la robótica. Considérense la obra de teatro de 1920 *R.U.R.*, que trata sobre robots esclavos que se rebelan contra sus dueños, la ya mencionada *2001: Una odisea en el espacio* de 1968, en la que una IA comienza a matar a la tripulación para conseguir llevar a cabo su misión, o la película de 2015 *Ex Machina*, en la que el robot de IA Ava se vuelve contra su creador. Las películas de *Terminator* también entran en esta narrativa de máquinas que se vuelven contra nosotros. El escritor de ciencia ficción Isaac Asimov denominó a este miedo «el complejo de Frankenstein»: el miedo a los robots. Esto es también relevante en la IA de hoy en día. Es algo con lo que tienen que lidiar científicos e inversores. Algunos argumentan contra ello; otros ayudan a crear y a mantener el miedo. Ya he mencionado a Musk. Otro ejemplo de una figura influyente que ha propagado el miedo a la IA es el físico Stephen Hawking, quien dijo en 2017 que la creación de IA podría ser el peor acontecimiento en la historia de la civilización (Kharpal 2017). El complejo de Frankenstein está generalizado y profundamente arraigado en la cultura y civilización occidentales.

La trascendencia y el apocalipsis de la IA

Ideas como el transhumanismo y la singularidad tecnológica tienen precedentes o, al menos, paralelismos en la historia de las religiones occidentales y en el pensamiento filosófico, especialmente en la tradición judeocristiana y en el platonismo. En contra de lo que mucha gente piensa, la religión y la tecnología siempre han estado conectadas en la historia de la cultura occidental. Limitaré mi discusión a la trascendencia y al apocalipsis.

En las religiones teístas, la trascendencia significa que un dios está «por encima» y es independiente del mundo material y físico, en oposición a estar en el mundo y ser parte del mundo (inmanencia). En la tradición mono-

teísta judeocristiana, Dios se considera que trasciende a su creación. Dios también puede ser visto al mismo tiempo como impregnando toda la creación y los seres (inmanencia) y, por ejemplo, en la teología católica Dios se entiende como revelándose a sí mismo inmanentemente a través de su hijo (Cristo) y el Espíritu Santo. Las narrativas frankensteinianas sobre la IA parecen subrayar la trascendencia en el sentido de una ruptura o brecha entre el creador y la creación (entre el *Homo deus* y la IA), sin dar demasiada esperanza a que pueda tenderse un puente que salve esta ruptura o brecha.

En contra de lo que mucha gente piensa, la religión y la tecnología siempre han estado conectadas en la historia de la cultura occidental.

La trascendencia también se puede referir a ir más allá de los límites, sobrepasando algo. En la historia de las religiones y en la filosofía de Occidente, esta idea a menudo toma la forma de ir más allá de los límites del mundo material y físico.

Por ejemplo, en el mundo mediterráneo del siglo II d.C., el gnosticismo entendía que toda la materia es mal y buscaba la liberación de la chispa divina del cuerpo humano. Anteriormente, Platón entendía el cuerpo como la cárcel del alma. En contraste con el cuerpo, el alma es vista como inmortal. En su metafísica, distingue entre las formas, que son eternas, y las cosas del mundo, que son cambiantes: las primeras trascienden así a las últimas. En el transhumanismo aparecen algunas ideas que recuerdan a esto. No solo se mantiene el fin de la trascendencia en el sentido de superar las limitaciones humanas, sino que las formas específicas en que se supone que ocurre esta trascendencia recuerdan a Platón y al gnosticismo: para alcanzar la inmortalidad, el cuerpo biológico debe trascenderse mediante la carga y el desarrollo de agentes artificiales. De forma más general, cuando la IA y la tecnología y la ciencia relacionada usan las matemáticas para extraer las más puras formas a partir del desordenado mundo material, tal actividad puede interpretarse como un programa platónico llevado a cabo por mediación de la tecnología. El algoritmo de la IA resulta ser una máquina platónica que extrae una forma (un modelo) del mundo de las apariencias (datos).

La trascendencia también puede significar sobrepasar la condición humana. En la tradición cristiana, esta puede constituirse como un intento de crear un puente entre Dios y los humanos convirtiendo a los humanos en dioses, quizás devolviéndoles su parecido original con Dios y su perfección (Noble 1997). Pero la búsqueda transhumanista de la inmortalidad es antigua. Se puede encontrar ya en la mitología mesopotámica: uno de los relatos escritos más antiguos de la humanidad, la *Epopeya de Gilgamesh,* cuenta la historia del rey de Uruk (Gilgamesh), quien busca la inmortalidad tras la muerte de su amigo Enkidu. No la encuentra; consigue recoger una planta que, según se dice, restaura la salud, pero una serpiente se la roba y, al final, tiene que aprender la lección de que debe enfrentarse a la realidad de su

propia muerte: la búsqueda de la inmortalidad es fútil. A lo largo de la historia de la humanidad, la gente ha buscado el elixir de la vida. La ciencia actual busca terapias anti-edad. En este sentido, la búsqueda transhumanista de la inmortalidad o la longevidad no es nada nuevo ni exótico: es uno de los sueños más antiguos de la humanidad y un exponente de parte de la ciencia contemporánea. En manos de los transhumanistas, la IA se convierte en una máquina trascendente que promete la inmortalidad.

Otros conceptos antiguos que pueden ayudarnos a contextualizar las ideas transhumanistas, y en particular la de la singularidad tecnológica: son el apocalipsis y la escatología. El término en griego antiguo *apocalypsis,* que también desempeña un papel en el mundo judío y cristiano, se refiere a la revelación. Hoy en día se refiere a menudo al contenido de un género particular de revelación: la visión de un fin de los tiempos o de un escenario de fin del mundo. En contextos religiosos, encontramos el término *escatología*: una parte de la teología preocupada por los acontecimientos finales de la historia y por el destino final de la humanidad. La mayoría de las ideas apocalípticas y escatológicas implican una radical y a menudo violenta disrupción o destrucción del mundo, en dirección hacia una realidad, ser y nivel de consciencia nuevos y más elevados. Esto también nos recuerda a los llamados cultos y sectas del juicio final, que trataban y tratan de predecir el desastre y el fin del mundo. Aunque, por norma general, los transhumanistas no tienen nada que ver con tales cultos y prácticas religiosas, claramente la idea de una singularidad tecnológica conlleva algún parecido con las narrativas apocalípticas, escatológicas y del juicio final.

Así, aunque el desarrollo de la IA esté basado en la ciencia, por lo cual se supone que no es un asunto ficción y se presenta secularizado, y aunque los transhumanistas tienden a distanciarse de la religión y rechazan cualquier insinuación de que sus trabajos son ficción, la ciencia ficción, las religiones antiguas y las ideas filosóficas entran inevitablemente en juego cuando discutimos en estos términos el futuro de la IA.

Cómo superar las narrativas competitivas y la exageración

Ahora bien, uno podría preguntarse: ¿existe una solución? ¿Podemos superar las narrativas competitivas y hallar nuevas formas de buscar un sentido inherente a la IA y otras tecnologías similares? ¿O da por hecho Occidente que la IA está condenada a permanecer en la prisión de estos miedos y fascinaciones modernos y sus antiguas raíces? ¿Podemos superar la exageración, o la discusión se mantendrá centrada en la superinteligencia? Creo que tenemos soluciones.

En primer lugar, cabe mirar más allá de la cultura occidental para encontrar diferentes tipos de narrativas no frankensteinianas en torno a la tecnología y formas no platónicas de pensamiento. Por ejemplo, en Japón, donde la cultura tecnológica está más influenciada por una religión de la naturaleza que en Occidente, en particular por el shinto, y donde la cultura popular ha retratado a las máquinas como agentes que ayudan, es habitual encontrar una actitud más amistosa hacia los robots y la IA. Allí no encontramos el complejo de Frankenstein. Lo que a veces se denomina una forma «animista» de pensamiento implica que las IAs pueden también en principio tener espíritu o alma, y también experimentarse como sagradas. Esto significa que no existe una narrativa competitiva; y ningún deseo platónico por transcender la materialidad y por defender constantemente al humano como algo que está por encima y más allá de la máquina, o que es fundamentalmente diferente de la máquina. Hasta donde yo sé, la cultura asiática tampoco tiene ideas acerca del fin de los tiempos. En contraste con las religiones monoteístas, las religiones de la naturaleza tienen una comprensión cíclica del tiempo. Así, mirar más allá de la cultura occidental (o, ciertamente, hacia nuestro propio pasado occidental, donde también encontramos religiones de la naturaleza) puede ayudarnos a evaluar críticamente las narrativas dominantes sobre el futuro de la IA.

Segundo, para ir más allá de la exageración y no limitar la discusión ética sobre la IA a los sueños y pesadillas proyectados sobre un futuro lejano, podemos (1) usar la filosofía y la ciencia para examinar y discutir críticamente los supuestos en torno a la IA y lo humano que desempeñan un papel en estos escenarios y discusiones (por ejemplo, ¿es posible la inteligencia artificial general? ¿Cuál es la diferencia entre humanos y máquinas? ¿Cuál es la relación entre los humanos y la tecnología? ¿Cuál es el estatus moral

de la IA?); (2) observar con más detalle la naturaleza de la IA existente y qué funciones cumple hoy día en sus diversas aplicaciones; (3) discutir los problemas éticos y sociales más concretos y acuciantes que plantea la IA tal y como está siendo aplicada hoy; (4) investigar las políticas de IA para el futuro próximo; y (5) preguntarnos si la atención que concentra el discurso público actual en la IA es útil a la luz de otros problemas a los que nos enfrentamos, y si la inteligencia debería ser nuestro único foco de atención. Seguiremos estos caminos en los siguientes capítulos.

Aunque, por norma general, los transhumanistas no tienen nada que ver con cultos y prácticas religiosas, claramente la idea de una singularidad tecnológica presenta ciertas similitudes con las narrativas apocalípticas, escatológicas y del juicio final.

Capítulo 3

Todo sobre los humanos

¿Es posible la IA general? ¿Existen diferencias fundamentales entre humanos y máquinas?

La visión transhumanista del futuro tecnológico supone que la inteligencia artificial general (o IA fuerte) es posible, ¿pero lo es? Esto es, ¿podemos crear máquinas con capacidades cognitivas similares a las humanas? Si la respuesta es no, entonces toda la visión de la superinteligencia es irrelevante para la ética de la IA. Si la inteligencia general humana no es posible en máquinas, no tenemos que preocuparnos por la superinteligencia. De forma más general, nuestra evaluación de la IA parece depender de lo que creamos que es la IA y en lo que puede convertirse, y de lo que pensemos sobre las diferencias entre humanos y máquinas. Al menos desde la mitad del siglo xx, los filósofos y los científicos han debatido sobre lo que los ordenadores pueden hacer y llegar a ser, y cuáles son las diferencias entre los humanos y las máquinas inteligentes. Echemos un vistazo a alguna de estas discusiones, que tratan tanto sobre lo que es y debería ser *el ser humano* como sobre lo que es y debería ser la IA.

¿Pueden tener los ordenadores inteligencia, consciencia y creatividad? ¿Pueden otorgar sentido a las cosas y entender significados? Existe toda una historia de crítica y escepticismo en torno a la posibilidad de una IA similar a la inteligencia humana. En 1972, Hubert Dreyfus, un filósofo con experiencia en fenomenología, publicó un libro titulado *What Computers Can't Do (Lo que los ordenadores no pueden hacer)*. Desde los años 60, Dreyfus ha sido muy crítico con todas las bases filosóficas de la IA y ha cuestionado sus promesas: sostuvo que el programa de investigación de la IA estaba destinado al fracaso. Antes de trasladarse a Berkeley, estuvo trabajando en

el MIT, un lugar importante para el desarrollo de IA, que, por entonces, estaba principalmente basada en la manipulación simbólica. Dreyfus argumentaba que el cerebro no es un ordenador y que la mente no opera mediante manipulación simbólica. Disponemos de un bagaje inconsciente de conocimiento común basado en la experiencia y en lo que Heidegger llamaría nuestro «ser-en-el-mundo», y este conocimiento es tácito y no puede formalizarse. Las destrezas humanas, decía Dreyfus, están basadas en el *know-how* (saber cómo) en vez de en el *know-that* (saber que). La IA no puede capturar este bagaje de significado y conocimiento; si ese es el objetivo de la IA, entonces es básicamente alquimia y mitología. Solo los seres humanos pueden discernir lo que es relevante porque, como seres encarnados y existenciales, estamos implicados en el mundo y somos capaces de responder a las demandas de la situación.

En aquel momento, Dreyfus encontró una fuerte oposición, pero más tarde muchos investigadores de la IA dejaron de prometer o predecir la IA general. La investigación en IA abandonó la dependencia de la manipulación simbólica en dirección hacia nuevos modelos, incluyendo aprendizaje automático basado en estadística. Y, al contrario que en la época de Dreyfus, en la que aún existía una gran separación entre la fenomenología y la IA, en la actualidad muchos investigadores de la IA adoptan enfoques científicos cognitivos encarnados y situados, que consideran más cercanos a la fenomenología.

Dicho esto, las objeciones de Dreyfus todavía son relevantes y demuestran que otras visiones de los seres humanos —especialmente, aunque no solo, en la llamada filosofía continental— chocan a menudo con las visiones científicas. La filosofía continental suele subrayar que los seres humanos y sus mentes son fundamentalmente distintos de las máquinas, y se centra en la experiencia de la (auto-)consciencia y en la existencia humanas, que no pueden ni deben ser reducidas a descripciones formales y explicaciones científicas. Otros filósofos, sin embargo, a menudo seguidores de la tradición analítica de la filosofía, respaldan una visión del ser humano que apoya a los investigadores de la IA que piensan que el cerebro y la mente *son y funcionan realmente* como sus equivalentes computacionales. Filósofos tales como Paul Churchland y Daniel Dennett son buenos ejemplos de

esto último. Churchland piensa que la ciencia, y en particular la biología evolutiva, la neurociencia y la IA, puede explicar completamente la consciencia humana. Cree que el cerebro es una red neuronal recurrente. Lo que él llama materialismo eliminativo niega la existencia de pensamientos y experiencias inmateriales. Lo que llamamos pensamientos y experiencias son simplemente estados mentales. Dennett niega también la existencia de cualquier cosa más allá de lo que ocurre dentro del cuerpo: cree que nosotros mismos «somos algo así como un robot» (Dennett 1997). Y si el ser humano es básicamente una máquina consciente, entonces tales máquinas son posibles, y no solamente en principio sino como cuestión de hecho. Podemos intentar crearlas. Curiosamente, tanto los filósofos continentales como los analíticos argumentan contra el dualismo cartesiano que separa mente y cuerpo, pero por diferentes motivos: los primeros porque creen que la existencia humana trata sobre ser-en-el-mundo, donde mente y cuerpo no están separados, mientras que los segundos, por razones materiales, opinan que la mente no es nada separada del cuerpo.

Existe toda una historia de crítica y escepticismo en torno a la posibilidad de una IA similar a la inteligencia humana.

Pero no todos los filósofos de la tradición analítica piensan que la IA general o fuerte sea posible. Desde el punto de vista del segundo Wittgenstein, se puede argüir que, si bien es cierto que un conjunto de leyes puede describir un fenómeno cognitivo, esto no implica que realmente tengamos tales leyes en la cabeza (Arkoudas y Bringsjord 2014). Como en la crítica de Dreyfus, esto cuestiona al menos *un tipo de IA,* la IA simbólica, siempre que se suponga que así es como piensan los seres humanos. Otra famosa crítica filosófica a la IA proviene de John Searle, quien argumenta contra la idea de que los programas de ordenador puedan tener estados cognitivos genuinos o comprender significados (Searle 1980). El experimento mental que ofrece, denominado «argumento de la habitación china», plantea lo siguiente: Searle está encerrado en una habitación y se le dan textos en chino, pero él no sabe chino. Sin embargo, puede responder a las preguntas que se le hacen en chino por hablantes de chino que están fuera de la habitación, ya que usa un libro de reglas que le permite producir las respuestas correctas *(output)* basándose en los documentos *(input)* que se le han dado. Puede hacerlo con éxito sin entender chino. De forma similar, dice Searle, los programas de ordenador pueden producir un *output* basado en un *input* mediante reglas que se les han dado, pero no entienden nada. En términos filosóficos más técnicos: los programas de ordenador carecen de intencionalidad, y el entendimiento genuino no puede generarse de forma computacional. Como Boden (2016) señala, la idea es que el significado proviene de los seres humanos.

A pesar de que hoy día los programas de ordenador de IA a menudo son diferentes de aquellos que criticaban Dreyfus y Searle, el debate continúa. Muchos filósofos piensan que existen diferencias importantes entre las formas en que los seres humanos y las máquinas piensan. Por ejemplo, todavía hoy se puede objetar que seamos seres vivos encarnados, conscientes y creadores de significado, cuya naturaleza, mente y conocimiento no pueden explicarse mediante la comparación con máquinas. Obsérvese, sin embargo, que incluso aquellos científicos y filósofos que creen que *en principio* existe una notable similaridad entre humanos y máquinas, y que *en teoría* la IA

general es posible, a menudo rechazan la visión de Bostrom de la superinteligencia e ideas parecidas que sostienen que una IA similar a un ser humano está a la vuelta de la esquina. Tanto Boden como Dennett creen que es muy difícil alcanzar la IA general en la práctica y que, por tanto, no es algo de lo que haya que preocuparse actualmente.

En el trasfondo de la discusión sobre la IA hay entonces grandes desacuerdos sobre la naturaleza del ser humano, la inteligencia humana, la mente, el entendimiento, la consciencia, la creatividad, el significado, el conocimiento humano, la ciencia, y así sucesivamente. Si de verdad se trata de una «batalla», esta concierne tanto a lo humano como a la IA.

Modernidad, (post)humanismo y postfenomenología

Es interesante, desde un punto de vista más amplio que el de las humanidades, contextualizar adicionalmente estos debates sobre la IA y el ser humano para entender qué es lo que está en juego. No solo tratan sobre tecnología y el ser humano, sino que reflejan profundas divisiones en la modernidad. Permítanme que me refiera brevemente a tres líneas divisorias que afectan indirectamente a las discusiones en el campo de la ética sobre la IA. La primera es la temprana división de la edad moderna entre la Ilustración y el Romanticismo. Las otras se desarrollan en época relativamente reciente: una se da entre el humanismo y el transhumanismo, que se sitúa dentro de las tensiones de la modernidad, y la otra entre el humanismo y el posthumanismo, que intenta ir más allá de la modernidad.

Una primera forma de hacer que el debate sobre la IA y el ser humano cobre sentido es considerar la tensión que existe en la modernidad entre la *Ilustración* y el *Romanticismo.* En los siglos XVIII y XIX, los pensadores ilustrados y los científicos desafiaron las visiones religiosas tradicionales y argumentaron que la razón, el escepticismo y la ciencia nos mostrarían a los humanos cómo es realmente el mundo, en oposición a cómo debería ser según ciertas creencias sin justificaciones argumentales y sin apoyo de la evidencia. Eran optimistas en cuanto a lo que la ciencia podría hacer para beneficiar a la humanidad. En respuesta, los románticos arguyeron que la razón abstracta y la ciencia moderna habían desencantado al mundo y que ne-

cesitábamos volver al misterio y la sorpresa que la ciencia pretendía eliminar. Observando el debate sobre la IA, parece que no nos hemos movido demasiado desde entonces. El trabajo de Dennett sobre la consciencia y el de Boden sobre la creatividad, por ejemplo, intentan deshacerse de ese misterio, de «romper el encantamiento», según explica Dennett. Estos pensadores son optimistas en tanto que creen que la ciencia puede desentrañar los misterios de la consciencia, de la creatividad, *etc.* Reaccionan contra aquellos que se resisten a los esfuerzos para *des-encantar* al ser humano, tales como los filósofos continentales que trabajan en la tradición del postmodernismo y subrayan el misterio del ser humano, o, en otras palabras, los nuevos románticos. «Romper el encantamiento, o mantener las incógnitas del ser humano» parece, entonces, una dicotomía sobre la que pivota la discusión acerca de la IA general y su futuro.

Una segunda tensión es la que se da entre los *humanistas* y los *transhumanistas*. ¿Qué es «el ser humano», y qué debería llegar a ser? ¿Es importante defender al ser humano como es, o debemos revisar el concepto que tenemos de él? Los humanistas celebran al ser humano tal como es. Hablando en términos éticos, subrayan el valor intrínseco y superior de los seres humanos. En el debate que rodea a la IA, se pueden encontrar trazas del humanismo en los argumentos que defienden los derechos humanos y la dignidad humana como bases de una ética de la inteligencia artificial, o en el argumento a favor de la centralidad de los seres humanos y sus valores en el desarrollo y el futuro de la IA. Aquí, el humanismo a menudo se coaliga con el pensamiento ilustrado, pero también puede tomar formas más conservadoras o románticas. El humanismo se puede encontrar también en la resistencia contra el proyecto transhumanista. Mientras que los transhumanistas piensan que deberíamos encaminarnos hacia un nuevo tipo de ser humano que esté mejorado mediante la ciencia y la tecnología, los humanistas defienden al ser humano tal y como es y subrayan su valor y dignidad, que se consideran amenazados por la filosofía y ciencia transhumanistas.

Somos seres vivos encarnados, conscientes y creadores de significado, cuya naturaleza, mente y conocimiento no pueden explicarse mediante la comparación con máquinas.

Las reacciones defensivas contra las nuevas tecnologías tienen su propia historia. En las humanidades y en las ciencias sociales, la tecnología ha sido a menudo criticada por amenazar a la humanidad y a la sociedad. Muchos filósofos del siglo XX, por ejemplo, eran muy pesimistas en cuanto a la ciencia y advertían contra una sociedad dominada por la tecnología. Pero ahora la batalla no es solo contra las vidas y la sociedad humanas, es contra el ser humano mismo: mejorarlo o no mejorarlo, esa es la cuestión. Por un lado, el ser humano en sí mismo se convierte en un proyecto científico-tecnológico, abierto a la mejora. Una vez se ha roto el encantamiento del ser humano (por Darwin, la neurociencia y la IA) podemos proseguir con su mejora. La IA nos ayuda a mejorar al ser humano. Por otro lado, deberíamos abrazar al ser humano tal y como es. Y algunos dirán: lo que es el ser humano siempre se nos escapa; no puede ser completamente comprendido por la ciencia.

Estas tensiones continúan dividiendo las mentes y corazones en torno a este debate. ¿Podemos superarlas? En la práctica, uno puede desechar la meta de crear una IA similar a un ser humano. Pero, incluso entonces, se mantienen los desacuerdos acerca del estatus de los *modelos de IA según modelos humanos* que utiliza la ciencia de la IA. ¿De veras nos enseñan algo acerca de la forma en que piensan los humanos? ¿O solo ilustran una determinada forma de pensar?, una que puede ser matemáticamente formalizada, por ejemplo, o una que busca el control y la manipulación. ¿Cuánto podemos aprender realmente sobre lo humano de estas tecnologías? ¿Es la humanidad más de lo que la ciencia puede abarcar? Incluso en las discusiones más moderadas emergen las luchas en torno a la modernidad.

Para encontrar una salida a este *impasse,* se puede seguir a los académicos de las humanidades y ciencias sociales que durante los últimos cincuenta años han explorado formas *no modernas* de pensamiento. Autores como Bruno Latour y Tim Ingold han evidenciado que es posible encontrar formas menos dualistas, más no-modernas de concebir el mundo que van más allá de la oposición Ilustración-Romanticismo. Podemos intentar cruzar la división moderna entre humanos y no humanos, no a través de la ciencia moderna o del transhumanismo que, a su manera, también conciben a los

humanos y a las máquinas como no fundamentalmente opuestos, sino a través del pensamiento posthumanista de las (post)humanidades. Esto nos lleva a la tercera tensión: entre *humanismo* y *posthumanismo*. Contra los humanistas, que son acusados de haber violentado a los no humanos tales como los animales en nombre del valor supremo de los seres humanos, los posthumanistas critican la centralidad del ser humano en las ontologías y en la ética modernas. De acuerdo con estos, los no humanos también importan, y no debemos tener miedo de cruzar las fronteras entre humanos y no humanos. Esta es una dirección interesante de explorar, ya que nos lleva más allá de la narrativa competitiva de los humanos y las máquinas.

Posthumanistas tales como Donna Haraway ofrecen una visión en la que convivir con las máquinas, e incluso fusionarse con ellas, no se percibe como una amenaza o una pesadilla, como en el humanismo, ni como el sueño humanista hecho realidad, sino de una manera en la que las fronteras ontológicas y políticas entre humanos y no humanos pueden y deben cruzarse. La IA es susceptible, entonces, de ser parte no del transhumanismo, sino de un proyecto posthumanista crítico, que entra en juego por el lado de las humanidades y las artes más que desde la ciencia. Las fronteras no se cruzan en nombre de la ciencia o del progreso universal, como algunos transhumanistas ilustrados querrían decir, sino en nombre de una política e ideología posthumanistas que estriba en cruzar fronteras. Y el posthumanismo puede también ofrecer algo más relevante para la IA: puede animarnos a reconocer que *los no humanos no necesitan ser similares a nosotros y no deben hacerse similares a nosotros*. Apoyada por un posthumanismo de este tipo, parece entonces que la IA puede liberarse de la carga de imitar o reconstruir lo humano y explorar formas diferentes de géneros y seres no humanos, de inteligencia, de creatividad, *etc.* La IA no necesita hacerse a nuestra imagen. El progreso significa aquí ir más allá de lo humano y abrirnos a nosotros mismos hasta lo humano para aprender de ello. Además, tanto los transhumanistas como los posthumanistas pueden aceptar que, en lugar de *competir* con la IA en una determinada tarea, podemos también establecer una meta común, que se alcanzaría entonces *colaborando* y mobilizando a los mejores agentes humanos y artificiales disponibles para así acercarnos a lograr esa meta común.

Apoyada por un posthumanismo de este tipo, parece entonces que la IA puede liberarse de la carga de imitar o reconstruir lo humano y explorar otras formas diferentes de géneros de seres no humanos, de inteligencia, de creatividad, *etc*.

Otra forma de dejar atrás la narrativa competitiva, que a veces se acerca al posthumanismo, es la de un enfoque propio de la filosofía de la tecnología denominado *postfenomenología.* Dreyfus recurre a la fenomenología, en particular al trabajo de Heidegger, pero el pensamiento postfenomenológico, iniciado por el filósofo Don Ihde, va más allá de la fenomenología de la tecnología *à la* Heidegger al centrarse en cómo se relacionan los seres humanos con tecnologías específicas y con artefactos materiales concretos. Dicho enfoque, que a menudo colabora con los estudios de la ciencia y la tecnología, nos recuerda la dimensión material de la IA. La IA es a menudo vista como poseedora de una naturaleza meramente abstracta o formal, ajena a los artefactos materiales e infraestructuras específicos, pero todas las formalizaciones, abstracciones y manipulaciones simbólicas mencionadas dependen de dichos instrumentos e infraestructuras. Por ejemplo, como veremos en el capítulo siguiente, la IA contemporánea depende en alto grado de las redes y de la producción de grandes cantidades de datos por medio de aparatos electrónicos. Estas redes y aparatos no son únicamente «virtuales» sino que tienen que producirse y prolongarse materialmente. Además, contra la moderna división sujeto-objeto, los postfenomenólogos como Peter-Paul Verbeek hablan de la constitución mutua de los humanos y la tecnología, sujeto y objeto. En lugar de plantearse la tecnología como una amenaza, subrayan que los seres humanos somos tecnológicos; esto es: siempre hemos usado la tecnología, es parte de nuestra existencia más que una cuestión externa que amenaza nuestra existencia; y esta tecnología, naturalmente, media en nuestra involucración en el mundo. Para la IA, esta visión parece implicar que la batalla humanista por defender al ser humano de la tecnología está mal enfocada. En su lugar, de acuerdo con esta perspectiva, el ser humano ha sido siempre tecnológico y, por tanto, deberíamos preguntarnos *cómo* media la IA en la relación de los seres humanos con el mundo e intentar moldear activamente estas mediaciones mientras nos sea posible: podemos y debemos discutir sobre la ética en la fase de desarrollo de la IA en lugar que quejarnos después de los problemas que causa.

Sin embargo, uno se podría preocupar por el hecho de que las visiones posthumanistas y postfenomenológicas no son lo suficientemente críticas, ya que son demasiado optimistas y están demasiado alejadas de la práctica científica y de la ingeniería, y no son lo bastante sensibles a los peligros reales y a las consecuencias éticas y sociales de la IA. Cruzar fronteras nunca antes cruzadas no es algo necesariamente exento de problemas y, en la práctica, tales ideas posthumanistas y postfenomenológicas pueden ser de poca ayuda ante la dominación y la explotación que puede que afrontemos como consecuencias de tecnologías como la IA. Uno también podría defender una visión más tradicional del ser humano o propugnar un nuevo tipo de humanismo, en lugar de un posthumanismo. De este modo, el debate continúa.

Capítulo 4

¿Simplemente máquinas?

Cuestionando el estatus moral de la IA: agencia y paciencia morales

Una de las cuestiones que aparecieron en el capítulo anterior era la de si los no humanos también importaban. Actualmente mucha gente piensa que los animales importan, moralmente hablando; pero esto no siempre fue así. Parece que antaño nos equivocamos con respecto a los animales. Si hoy en día mucha gente pensase que las IAs son simplemente máquinas, ¿estaría cometiendo un error similar? ¿Merecen las IAs superinteligentes que se les reconozca, por ejemplo, un estatus moral? ¿Deberían otorgárseles derechos? ¿O es una idea peligrosa el solo hecho de considerar si las máquinas pueden tener un estatus moral?

Una forma de discutir lo que es la IA y lo que puede llegar a ser es revisar el estatus moral de la IA. Aquí planteamos cuestiones filosóficas que atañen a la IA, no a través de la metafísica, la epistemología o la historia de las ideas, sino a través de la filosofía moral. El término *estatus moral* (también denominado a veces *rango moral)* se puede referir a dos tipos de cuestiones. La primera tiene que ver con lo que la IA es capaz de hacer moralmente hablando: en otras palabras, si pueden tener lo que los filósofos denominan *agencia moral* y, si es así, si puede ser un agente moral completo. ¿Qué significa esto? En la actualidad, los actos de las IAs ya tienen consecuencias morales. La mayoría de la agente estará de acuerdo en que la IA tiene una forma «débil» de agencia moral en este sentido, que es similar a, por ejemplo, la mayoría de los coches actuales: estos últimos también pueden generar problemas de índole moral. Pero dado que la IA se está volviendo cada vez más inteligente y autónoma, ¿puede tener una forma más fuerte de agencia moral? ¿Se le debería otorgar, o desarrollará la IA, alguna

capacidad para el razonamiento moral, el juicio y la toma de decisiones? Por ejemplo: ¿pueden y deben los coches autónomos que usan IA ser considerados agentes morales? Estas cuestiones atañen a la ética de la IA, pues plantean qué tipo de capacidades morales tiene o debería tener una IA. Pero las cuestiones sobre el «estatus moral» pueden también referirse a cómo debemos tratar a una IA. ¿Es una IA «una simple máquina», o merece alguna forma de consideración moral? ¿Deberíamos tratarla de forma diferente a, por ejemplo, una tostadora o una lavadora? ¿Deberíamos otorgar derechos a una entidad artificial altamente inteligente, si tal entidad se desarrollase algún día, incluso si no fuese humana? Esto es lo que los filósofos denominan el problema de la *paciencia moral,* que no gira en torno a la ética *de* o *en* la IA sino a *nuestra* ética para *con* la IA. Aquí, la IA es el objeto de la preocupación ética, no un agente ético potencial en sí mismo.

¿Es una IA «una simple máquina», o merece alguna forma de consideración moral? ¿Deberíamos tratarla de forma diferente a, por ejemplo, una tostadora o una lavadora?

Agencia moral

Comencemos con la cuestión de la agencia moral. Si existiese una IA mucho más inteligente que las que existen en la actualidad, podemos suponer que podría desarrollar un razonamiento moral y que podría aprender cómo toman decisiones los seres humanos sobre problemas éticos. ¿Pero, sería esto suficiente para otorgarle una agencia moral completa, esto es, una agencia moral similar a la humana? La cuestión no incumbe solo a la ciencia ficción. Si ya delegamos algunas de nuestras decisiones a algoritmos, por ejemplo en los coches o en los juzgados, sería positivo que estas decisiones fuesen moralmente correctas. Pero no está claro que las máquinas tengan las mismas capacidades morales que los seres humanos. Se les otorga agencia en el sentido de que hacen cosas en el mundo, y estas acciones tienen consecuencias morales. Por ejemplo, un coche autónomo puede causar un accidente, o una IA puede recomendar enviar a una persona concreta a la cárcel. Estos comportamientos y elecciones no son neutrales en términos de moral: claramente tienen consecuencias morales para la gente que está involucrada en dichos acontecimientos. ¿Pero debería, para lidiar con este problema, otorgarse agencia moral a las IAs? ¿Pueden tener agencia moral completa?

Hay distintas posturas filosóficas en torno a estas cuestiones. Algunos dicen que las máquinas jamás podrán ser agentes morales. Las máquinas, argumentan, no tienen las capacidades necesarias para la agencia moral, tales como estados mentales, emociones o libre albedrío. De ahí que sea peligroso suponer que puedan tomar decisiones morales correctas y delegar totalmente en ellas dichas decisiones. Por ejemplo, Deborah Johnson (2006) ha argumentado que los sistemas de ordenadores no tienen agencia moral por sí mismos: están producidos y son utilizados por humanos, y solo estos humanos tienen libertad y son capaces de actuar y decidir moralmente. De forma similar, se podría decir que las IAs están creadas por humanos y que, por tanto, la toma de decisiones morales en prácticas tecnológicas debería ser ejercida por humanos. En el otro extremo del espectro están aquellos

que piensan que las máquinas pueden ser agentes morales completos de la misma forma que lo son los seres humanos. Investigadores tales como Michael y Susan Anderson, por ejemplo, afirman que, en principio, es posible y deseable otorgar a las máquinas una moralidad de tipo humano (Anderson y Anderson 2011). Podemos dar principios morales a las IAs, y las máquinas pueden incluso ser mejores que los seres humanos en el razonamiento moral, ya que son más racionales y no se dejan llevar por sus emociones. Contra esta postura, algunos han argumentado que las reglas morales entran a menudo en conflicto (considérense, por ejemplo, las historias de robots de Asimov, en las que las leyes morales para los robots siempre meten en problemas a los robots y a los humanos) y que el proyecto completo de construir «máquinas morales» dándoles reglas se basa en suposiciones erróneas con respecto a la naturaleza de la moralidad. La moralidad no puede reducirse a seguir reglas y no es enteramente una cuestión de emociones humanas: pero estas pueden bien ser indispensables para el juicio moral. Si fuese posible la IA general, resultaría indeseable una «IA psicópata» que fuera perfectamente racional pero insensible a las preocupaciones humanas porque carece de emociones (Coeckelbergh 2010).

Según estas razones, podría o bien rechazarse en su conjunto la idea misma de la agencia moral completa, o bien adoptarse una posición intermedia: tenemos que dar a las IAs algún tipo de moralidad, pero no moralidad completa. Wendell Wallach y Colin Allen usan el término «moralidad funcional» (2009, pág. 39). Los sistemas de IA necesitan alguna capacidad para evaluar las consecuencias éticas de sus acciones. La razón de esta decisión está clara en el caso de los coches autónomos: el coche probablemente estará involucrado en situaciones donde haya que tomar una decisión moral, pero no hay tiempo suficiente para una toma de decisión o intervención humana. A veces estas elecciones se presentan como dilemas. Los filósofos hablan del *dilema del tranvía,* llamado así a partir del experimento mental en el que un tranvía avanza sin control por unos raíles y alguien tiene que escoger entre no hacer nada, con lo que morirán cinco personas que están atadas a la vía, o tirar de una palanca que hace que el tranvía siga por otro raíl en el que solo hay atada una persona, a la que no se conoce. ¿Cuál es la solución moralmente correcta? De forma similar, los que proponen este en-

foque argumentan que un coche autónomo puede tener que hacer la elección moral entre, por ejemplo, matar a unos peatones que cruzan la carretera y estrellarse contra un muro, matando, así, al conductor. ¿Qué debería escoger el coche? Parece que tendremos que tomar estas decisiones morales (de antemano) y asegurarnos de que los diseñadores las implementan en los coches. O, quizás, necesitemos construir coches con IAs que aprendan de las elecciones humanas. Sin embargo, se podría cuestionar si proporcionar reglas a las IAs es una buena forma de representar la moralidad humana, si es que la moralidad puede «representarse» y reproducirse, y si los dilemas del tranvía capturan de alguna forma algo que es central para la vida y la experiencia moral. O, desde una perspectiva totalmente diferente, se puede cuestionar si los humanos son, de hecho, buenos a la hora de tomar decisiones morales. ¿Por qué imitar en cualquier caso la moralidad humana? Los transhumanistas, por ejemplo, pueden argumentar que las IAs tendrán una moralidad superior porque serán más inteligentes que nosotros.

Este cuestionamiento del énfasis en lo humano nos lleva a otra posición que no necesita de una agencia moral completa, y que intenta alejarse de la posición ética antropocéntrica. Luciano Floridi y J. W. Sanders (2004) han apostado por una moralidad *a-mental,* que no se base en las características de los seres humanos. Podríamos hacer que la agencia moral dependiera de tener un nivel suficiente de interactividad, autonomía y adaptabilidad, así como de ser capaces de acción moralmente calificable. De acuerdo con estos criterios, un perro de rescate es un agente moral, pero también lo es un *bot* web con IA que filtra los emails no deseados. De forma similar, se podrían aplicar criterios no antropocéntricos para la agencia moral de los robots, como propuso John Sullins (2006): si una IA es autónoma respecto de los programadores y podemos explicar su comportamiento atribuyéndole intenciones morales (como la intención de hacer bien o mal), y si se comporta de una forma que muestra un entendimiento de su responsabilidad frente a otros agentes morales, entonces esta IA es un agente moral. Desde esta perspectiva, pues, las IAs no necesitan de una agencia moral completa en tanto que esto implica agencia moral humana, sino que definen la agencia moral de una forma que es, en principio, independiente de la agencia moral humana completa y de las capacidades humanas necesarias para ello.

Sin embargo, ¿sería suficiente tal agencia moral artificial si la juzgáramos desde los estándares morales humanos? El problema en la practica es que, por ejemplo, los coches autónomos pueden no ser lo bastante morales. La preocupación por lo que respecta a los principios es que aquí nos apartamos demasiado de la moralidad humana. Mucha gente piensa que la agencia moral es y debe estar vinculada a lo que entendemos por ser humano y ser persona. No están dispuestos a apoyar nociones posthumanistas o transhumanistas.

Paciencia moral

Otra controversia se refiere a la paciencia moral de la IA. Imaginemos que tenemos una IA superinteligente. ¿Es moralmente aceptable apagarla, «matarla»? Y algo más cercano a la IA actual: ¿está bien golpear a un robot con IA? . Si las IAs van a ser parte de la vida cotidiana, como muchos investigadores predicen, entonces estos casos aparecerán inevitablemente y harán surgir la cuestión de cómo nos deberíamos comportar los humanos con estas entidades artificiales. Una vez más, sin embargo, no es necesario que busquemos respuestas en un futuro lejano o en la ciencia ficción. La investigación ha demostrado que la gente empatiza con los robots y que duda a la hora de «matarlos» o «torturarlos» (Suzuki *et al.* 2015; Darling, Nandy y Breazeal 2015), incluso si estos robots no tienen IA. Los humanos parecen necesitar muy poco de los agentes artificiales para proyectar sobre ellos personalidad o humanidad, o para empatizar con ellos. Si estos agentes se convirtieran ahora en IA, lo cual los haría potencialmente más parecidos a los humanos (o a los animales), es posible que esta cuestión de la paciencia moral se hiciera más acuciante. Por ejemplo, ¿cómo deberíamos reaccionar ante las personas que empatizan con la IA? ¿Están equivocados?

Decir que las IAs son simples máquinas y que la gente que empatiza con ellas está equivocada en su juicio, emociones y experiencia moral es, quizás, la posición más intuitiva. A primera vista, parece que no tenemos obligación alguna con respecto a las máquinas. Son cosas, no personas. Muchos investigadores de la IA siguen esta línea de pensamiento. Por ejemplo, Joanna Bryson ha argumentado que los robots son herramientas, tienen pro-

pietarios y no tenemos obligaciones hacia ellos (Bryson 2010). Aquellos que mantienen esta posición puede que también estén de acuerdo en que en el caso de que las IAs llegaran a ser conscientes, a tener estados mentales, y demás, deberíamos otorgarles un estatus moral, aunque dirán que esta condición no se cumple actualmente. Como hemos visto en los capítulos anteriores, algunos argumentarán que dicha situación no se dará nunca; otros que podría darse en principio, pero que no ocurrirá en un lapso de tiempo cercano. En cualquier caso, su respuesta a la pregunta sobre el estatus moral es que hoy en día y en el futuro próximo las IAs tienen que ser tratadas como cosas, a menos que se demuestre que son lo contrario.

Sin embargo, un problema de esta posición es que no explica ni justifica nuestras intuiciones y las experiencias morales que nos dicen que hay *algo* que está mal cuando se «maltrata» a una IA, incluso si esta no tiene propiedades similares a las humanas o animales como la consciencia o la sensibilidad. Para encontrar tales justificaciones se podría acudir a Kant, que argumentó que está mal dispararle a un perro, no porque hacerlo incumpla ninguna obligación para con el perro, sino porque tal persona «daña la amabilidad y las cualidades humanas que posee en sí misma, que debería ejercitar en virtud de sus obligaciones para con la humanidad» (Kant 1997). Hoy en día tendemos a pensar en los perros de otro modo (aunque no todo el mundo ni en todos los sitios), pero parece que el argumento podría aplicarse a las IAs: podríamos decir que no tenemos obligaciones con respecto a las IAs, pero aun así no deberíamos patearlas o «torturarlas» porque ello nos hace crueles para con los humanos. Se podría usar también el argumento de la ética de la virtud, que es también indirecto, ya que versa sobre los humanos, no sobre la IA: «maltratar» a una IA está mal no porque se le haga ningún daño a la IA, sino porque en caso de hacerlo se daña nuestra integridad moral. No nos hace mejores personas. Contra este enfoque podemos argüir que en el futuro algunas IAs puedan tener un valor intrínseco y merezcan nuestra atención moral, suponiendo que tengan cualidades tales como la sensibilidad. Una obligación indirecta o un enfoque basado en la virtud no parece tomarse en serio esta «otra» cara de la relación moral. Solo se preocupa por los humanos. ¿Qué pasa con las IAs? ¿Pueden las IAs o los robots

ser *otros,* como ha preguntado David Gunkel (2018)? De nuevo, el sentido común apunta a que no: las IAs no reúnen las cualidades necesarias.

Hay quien considera que «maltratar» a una IA está mal no porque se le haga ningún daño a la IA, sino porque en caso de hacerlo se daña nuestra integridad moral.

Desde un enfoque totalmente distinto se argumenta que la manera en que nos cuestionamos el estatus moral es problemática. El razonamiento moral usual sobre el estatus moral está basado en cuáles son las propiedades moralmente relevantes de las entidades en cuestión: por ejemplo, consciencia o sensibilidad. ¿Pero cómo sabemos si la IA realmente reúne o no las cualidades moralmente relevantes? ¿Estamos seguros en el caso de *los humanos?* Los escépticos dirían que no estamos seguros. Incluso sin dicha certeza epistemológica atribuimos un estatus moral a los humanos sobre la base de la apariencia. Puede que esto sucediera si en el futuro las IAs tuvieran una apariencia y comportamiento similares a los de los humanos. Parece que, independientemente de lo que los filósofos consideren moralmente *correcto,* los humanos continuarán atribuyendo estatus moral a máquinas y, en consecuencia, les concederán derechos. Además, si analizamos más de cerca cómo los humanos otorgamos estatus moral *de facto,* resulta que, por ejemplo, las relaciones sociales existentes y el lenguaje desempeñan un notable papel. Por ejemplo, si tratamos a nuestro gato con cariño, no es porque llevemos a cabo un juicio moral sobre él, sino porque hemos adquirido un tipo de relación social con él. Ya es una mascota y un compañero antes de que realicemos la operación filosófica de otorgarle un estatus moral (si es que llegamos alguna vez a sentir dicha necesidad). Y, si le ponemos a nuestro perro un nombre, entonces (en contraste con los animales sin nombre que nos comemos), ya le hemos conferido un estatus moral particular independientemente de sus cualidades objetivas. Bajo un enfoque de este tipo, relacional, crítico y no dogmático (Coeckelbergh 2012), podríamos concluir, del mismo modo, que los seres humanos otorgaremos el estatus a las IAs y que este dependerá de cuán integradas estén ellas en nuestra vida social, en el lenguaje y en la cultura humana.

Además, dado que tales condiciones son susceptibles de cambiar a lo largo de la historia (pensemos de nuevo en cómo tratábamos a los animales y pensábamos en ellos), quizás fuera necesaria alguna precaución antes de «fijar» el estatus moral de la IA en general o de cualquier IA concreta. ¿Y por qué incluso hablar de la IA en general o en abstracto? Parece que hay

algo que no encaja en el proceso por el cual otorgamos estatus moral: para juzgarlo, tomamos en consideración la entidad fuera de su contexto relacional, y antes de que tengamos el resultado de nuestro procedimiento moral ya la tratamos, de forma jerárquica, condescendiente y hegemónica, como una entidad sobre la que tomamos decisiones los seres humanos como jueces superiores. Parece que antes de que llevemos a cabo el juicio sobre su estatus moral, ya nos hemos posicionado ante esa entidad y, quizás, incluso la hayamos violentado al tratarla como el objeto de nuestra toma de decisión, estableciéndonos a nosotros mismos como dioses de la Tierra, poderosos y omniscientes, que se reservan el derecho a conferir un estatus moral a otras entidades. También hacemos invisibles todos los contextos situacionales y sociales. Como en el caso del dilema del tranvía, hemos reducido la ética a una caricatura. Con tales razonamientos, los filósofos morales parecen hacer lo que los filósofos que siguen a Dreyfus acusan de hacer a los investigadores de la IA simbólica: formalizar y abstraer una gran fuente de experiencia moral y de conocimiento a costa de dejar fuera lo que nos hace humanos y, además, con el riesgo de pedir la cuestión misma del estatus moral de los no humanos. Independientemente de cuál «sea» el estatus moral actual de las IAs, si es que este puede acotarse por completo e independientemente de la subjetividad humana, merece la pena examinar con perspectiva crítica nuestra propia actitud moral y el proyecto del razonamiento moral abstracto mismo.

Hacia los problemas prácticos

Como demuestran las cuestiones examinadas en este capítulo y el previo, pensar sobre la IA no solo nos permite aprender cosas sobre ética en sí. También nos enseña cosas de nosotros mismos: de cómo pensamos y cómo nos relacionamos y deberíamos relacionarnos con los no humanos. Si estudiamos los fundamentos filosóficos de la ética de la IA, encontramos profundos desacuerdos sobre la naturaleza y el futuro de la humanidad, la ciencia y la modernidad. Cuestionar la IA abre un abismo de problemas críticos sobre el conocimiento, la sociedad y la naturaleza de la moralidad humanas.

Los debates filosóficos en torno a esto son menos rebuscados y menos «académicos» de lo que se podría pensar. Continuarán resurgiendo cuando consideremos más adelante los problemas éticos, legales y políticos concretos suscitados por la IA. Si intentamos abordar temas como la responsabilidad y los coches autónomos, la transparencia del aprendizaje automático, la IA sesgada, o la ética de los robots sexuales, pronto nos encontraremos de nuevo confrontados con ellos. Si la ética de la IA quiere ser más que una lista de problemas a tener en cuenta, debería también aportar algo sobre tales cuestiones.

Dicho esto, es hora de volver a las cuestiones prácticas. Estas no conciernen ni a los problemas filosóficos planteados por una hipotética inteligencia artificial general ni a los riesgos que lleva aparejados una superinteligencia en un futuro lejano, ni cualesquiera otros monstruos espectaculares de la ciencia ficción. Por el contrario, afectan a las realidades menos visibles y probablemente menos atractivas, pero aun así importantes, de las IAs ya implantadas. La IA tal como ya funciona actualmente no desempeña el papel del monstruo de Frankenstein o del espectacular robot que amenaza la civilización, y es más que un experimento mental filosófico. La IA concierne a las tecnologías menos visibles, ocultas pero dominantes, poderosas, y cada vez más inteligentes que ya conforman nuestras vidas a día de hoy. Las éticas de la IA se ocupan de los desafíos planteados por la IA actual y del futuro cercano y su impacto en nuestras sociedades y en las democracias vulnerables. La ética de la IA se ocupa de la vida de la gente y de la política. Se ocupa de nuestra necesidad, como personas y sociedades, de lidiar con los problemas éticos de *ahora.*

Capítulo 5

La tecnología

Antes de discutir problemas éticos de la IA de forma más detallada y concreta, tenemos otra tarea pendiente para despejar el camino: una vez superada la exageración, nos hace falta una mejor comprensión de la tecnología y sus aplicaciones. Dejando de lado la ciencia ficción transhumanista y la especulación filosófica sobre la IA general, echemos un vistazo a lo que es y hace la IA actualmente. Ya que las definiciones de la IA y otros términos son controvertidas en sí mismas, no ahondaré demasiado en los debates filosóficos o en la contextualización histórica. Mi principal propósito es ofrecer al lector una idea de la tecnología en cuestión y de cómo se usa. Déjenme comenzar diciendo algo sobre la IA en general; el siguiente capítulo se centra en el aprendizaje automático, la ciencia de datos y sus aplicaciones.

¿Qué es la Inteligencia Artificial?

La IA se puede definir como una inteligencia desplegada o simulada por un código (algoritmos) o por máquinas. Esta definición de la IA plantea el problema de cómo definir la inteligencia. Hablando filosóficamente, es un concepto vago. Una comparación obvia es la inteligencia de tipo humano. Por ejemplo, Philip Jansen *et al.* definen la IA como «la ciencia y la ingeniería de máquinas con capacidades que se consideran inteligentes según los estándares de la inteligencia humana» (2018, pág. 5). Desde esta perspectiva, la IA tiene como objeto el crear máquinas inteligentes que puedan pensar o (re)accionar como lo hacen los humanos. Sin embargo, muchos investigadores de la IA piensan que la inteligencia no necesita ser de corte humano y prefieren una definición más neutral que se formula en tér-

minos independientes de la inteligencia humana y de las metas relacionadas con la IA general o fuerte. Enumeran todo tipo de funciones cognitivas y tareas tales como aprendizaje, percepción, planificación, procesamiento del lenguaje natural, razonamiento, toma de decisiones y solución de problemas (la última a menudo se hace equivaler con la inteligencia *per se).* Por ejemplo, Margaret Boden declara que la IA «busca hacer que los ordenadores hagan el tipo de cosas que puede hacer la mente». A primera vista, esto parece sonar como si los humanos fueran el único modelo. Sin embargo, Boden enumera a continuación todo tipo de destrezas psicológicas como la percepción, la predicción y la planificación, que forman parte de los «entornos ricamente estructurados de distintas capacidades de procesamiento de la información» (2016, pág. 1). Y este procesamiento de la información no tiene por qué ser una cuestión exclusivamente humana. La inteligencia general, de acuerdo con Boden, no es humana por necesidad. Algunos animales también pueden ser considerados inteligentes. Y los transhumanistas sueñan con mentes futuras que ya no estén biológicamente integradas. Dicho esto, la meta de alcanzar capacidades similares a las humanas y, posiblemente, una inteligencia general de tipo humano ha sido parte de la IA desde el inicio.

La historia de la IA está estrechamente conectada con la informática y con disciplinas relacionadas como las matemáticas y la filosofía y, por tanto, se remonta al menos hasta el comienzo de la Edad Moderna (Gottfried Wilhelm Leibniz y René Descartes, por ejemplo) si no a la antigüedad, con sus historias sobre artesanos que creaban seres artificiales e ingeniosos artefactos mecánicos capaces de engañar a la gente (piensen en las figuras animadas de la antigua Grecia o en las figuras mecánicas con forma humana de la antigua China). Pero como disciplina propiamente dicha se considera que la IA comenzó en los años 50 del siglo pasado, tras la invención de los ordenadores programables en los años 40 y con el nacimiento de la disciplina de la cibernética, definida por Norbert Wiener en 1948 como el estudio científico del «control y comunicación en el animal y en la máquina» (Wiener 1948). Un momento importante en la historia de la IA fue la publicación en *Mind* en1950 del artículo de Alan Turing «Computing Machinery and Intelligence» [«Maquinaria computacional e inteligencia»], en el que intro-

dujo el famoso «test de Turing», aunque trataba más ampliamente la cuestión de si las máquinas pueden pensar y ya especulaba acerca de si las máquinas podrían aprender y llevar a cabo tareas abstractas. Pero el seminario Dartmouth que tuvo lugar en verano de 1956 en Hannover, New Hampshire, se considera generalmente el lugar de nacimiento de la IA contemporánea. Su organizador, John McCarthy, acuñó el término IA, y entre los participantes estaban incluidos los nombres de Marvin Minsky, Claude Shannon, Allen Newell y Herbert Simon. Dado que la cibernética parecía estar demasiado dedicada a las máquinas analógicas, la IA de Dartmouth se centró las máquinas digitales. La idea era *simular* la inteligencia humana (que no recrearla: el proceso no es el mismo que en los humanos). Muchos participantes pensaban que una máquina tan inteligente como un ser humano estaría a la vuelta de la esquina, y de hecho que no llevaría más que una generación.

Esta es la meta de la *IA fuerte.* La IA *fuerte* o *general* es capaz de llevar a cabo cualquier tarea cognitiva que puedan realizar los humanos, mientras que la IA *débil* o *estrecha* solo puede operar en ámbitos específicos como el ajedrez, la clasificación de imágenes, *etc.* Hoy por hoy, no hemos alcanzado la IA general y, como hemos visto en los capítulos anteriores, se duda de que jamás lleguemos a alcanzarla. A pesar de que algunos investigadores y empresas están intentando desarrollarla, especialmente aquellas que creen en la teoría computacional de la mente, la IA general no se vislumbra en el horizonte. De ahí que las cuestiones éticas y políticas del capítulo siguiente se centren en la IA débil o estrecha, que ya tenemos hoy en día y que es probable que se vuelva más poderosa y dominante en un futuro próximo.

La IA se puede definir o bien como una ciencia o bien como una *tecnología.* Su objetivo puede interpretarse como el de alcanzar una mejor explicación científica de la inteligencia y de las mencionadas funciones cognitivas. Puede ayudarnos a comprender mejor a los seres humanos y a otros seres dotados de inteligencia natural. En este sentido, es una ciencia y una disciplina que estudia sistemáticamente el fenómeno de la inteligencia (Jansen *et al.* 2018) y, a veces, la mente o el cerebro. Como tal, la IA está relacionada con otras ciencias como la ciencia cognitiva, la psicología, la ciencia de datos (véase más adelante) y, a veces, también la neurociencia, que llega a sus propias conclusiones acerca de la comprensión de la inteligencia natu-

ral. Pero la IA también puede buscar el desarrollo de tecnologías para diversos objetivos prácticos, o para «lograr cosas útiles», como dice Boden: puede darse en forma de herramientas, diseñadas por humanos, que generen una apariencia de inteligencia y de comportamiento inteligente con fines prácticos. Las IAs pueden hacer esto analizando el entorno (datos) y actuando con un grado importante de autonomía. A veces, los intereses científico-teóricos y los fines tecnológicos coinciden; esto sucede, por ejemplo, en la neurociencia computacional, que utiliza herramientas de la informática para comprender el sistema nervioso, o en proyectos particulares como el «Proyecto cerebro humano» europeo, que implica a la neurociencia pero también a la robótica y a la IA. Algunos de estos proyectos combinan neurociencia con aprendizaje automático y con la llamada neurociencia del *big data* (por ejemplo, Vu *et al.* 2018).

De forma más general, la IA se basa en y está relacionada con muchas disciplinas, incluyendo las matemáticas (por ejemplo, la estadística), la ingeniería, la lingüística, la ciencia cognitiva, la informática, la psicología e, incluso, la filosofía. Como hemos visto, tanto los filósofos como los investigadores de la IA están interesados en comprender la mente y fenómenos como la inteligencia, la consciencia, la percepción, la acción y la creatividad. La IA está influenciada por la filosofía y viceversa. Keith Frankish y William Ramsey reconocen esta conexión con la filosofía, subrayan la multidisciplinariedad de la IA y combinan los aspectos científicos y tecnológicos cuando definen la IA como «un enfoque multidisciplinar para la comprensión, modelado y reproducción de la inteligencia y los procesos cognitivos mediante el uso de varios principios y dispositivos computacionales, matemáticos, lógicos, mecánicos e, incluso, biológicos» (2014, 1). La IA es, pues, tanto teórica como práctica, tanto ciencia como tecnología. Este libro se centra en la IA como una tecnología, en el sentido más práctico: no solo porque dentro de la IA se ha desplazado el foco en esta dirección, sino especialmente porque es sobre todo en esta forma en la que la IA conlleva consecuencias éticas y sociales (a pesar de que la investigación científica tampoco es totalmente neutral en cuanto a la ética).

Como tecnología, la IA puede adoptar varias formas y normalmente es parte de sistemas tecnológicos más amplios: algoritmos, máquinas, robots,

etc. Por ello, aunque la IA puede tener que ver con «máquinas», este término no solo se refiere a los robots, y menos todavía a robots humanoides. La IA puede estar integrada en muchos otros tipos de sistemas y dispositivos tecnológicos. Los sistemas de IA pueden adoptar la forma de un *software* que funciona en la web (por ejemplo, *bots* de *chat*, motores de búsqueda, análisis de imagen), pero también puede estar integrada en aparatos físicos como robots, coches, o en aplicaciones del «internet de las cosas» . Para el internet de las cosas se usa a veces el término «sistemas ciberfísicos»: dispositivos que funcionan en, e interactúan con, el mundo físico. Los robots son un tipo de sistemas ciberfísicos que influyen directamente en el mundo (Lin, Abney y Bekey 2011).

Si la IA está integrada en un robot, a veces se dice que es una IA *corporeizada.* Al ejercer una influencia directa en el mundo físico, la robótica resulta altamente dependiente de sus componentes materiales. Pero cada IA, incluyendo el software activo en la web, «hace» algo y también tiene aspectos materiales, tales como el ordenador en el que se ejecuta, los aspectos materiales de la red y de la infraestructura de la que depende, y un largo etcétera. Esta cuestión hace que sea problemática la distinción entre, por una parte, las aplicaciones «virtuales» basadas en la web y el «software» y, por otra, las aplicaciones físicas o «hardware». El software de IA necesita infraestructura física y *hardware* para ejecutarse, y los sistemas ciberfísicos *son* «IA» solamente si están conectados al software pertinente. Además, fenomenológicamente hablando, el hardware y el software a veces se fusionan en nuestra experiencia y uso de dispositivos: no experimentamos un robot interactivo humanoide impulsado por IA, o un aparato de IA conversacional como Alexa, bien como software o bien como hardware, sino como un único dispositivo tecnológico (y, a veces, casi como una persona, como sucede, por poner un caso, con Hello Barbie).

Es probable que la IA ejerza una influencia importante en la robótica, por ejemplo, a través del progreso en el procesamiento de lenguajes naturales y en las comunicaciones que simulan a la humana. A menudo se les llama a estos robots «robots sociales» porque están diseñados para participar en la vida social ordinaria de los seres humanos, por ejemplo, interactuando, ya sea como acompañantes o como asistentes, con los humanos de for-

ma natural. Así, la IA puede fomentar un mayor desarrollo de la robótica social.

Sin embargo, independientemente de la apariencia y del comportamiento del sistema como un todo y su influencia en su entorno, que es muy importante fenomenológica y éticamente hablando, la base de la «inteligencia» de una IA es el software: un *algoritmo* o una combinación de algoritmos. Un algoritmo es un conjunto y secuencia de instrucciones, como una receta, que le dice qué hacer al ordenador, *smartphone,* máquina, robot, o cualquier cosa en la que esté integrado. Conduce a un resultado *(output)* particular basándose en la información disponible *(input).* Se utiliza para resolver un problema. Para entender la ética en la IA, necesitamos comprender cómo funcionan los algoritmos de IA y qué hacen. Diré más sobre este tema en este y en el siguiente capítulo.

Diferentes enfoques y subcampos

Existen diferentes tipos de IA. También se podría decir que hay diferentes *enfoques* o *paradigmas de investigación.* Como vimos a partir de la crítica de Dreyfus, por lo general, la IA ha sido fundamentalmente *simbólica.* Este fue el paradigma dominante hasta finales de los años 80. La IA simbólica se basa en las representaciones simbólicas de tareas cognitivas superiores tales como el razonamiento abstracto y la toma de decisiones. Por ejemplo, puede decidir basándose en un *árbol de decisiones* (un modelo de decisiones y sus posibles consecuencias, a menudo representado visualmente como un gráfico de flujo). Un algoritmo que hace esto contiene afirmaciones condicionales: reglas de decisión de forma *si* … (condiciones) … *entonces* … (resultado). El proceso es determinista. Mediante el empleo de una base de datos que representa el conocimiento humano experto, una IA puede razonar utilizando una gran cantidad de información y actuar como un *sistema experto.* Puede tomar decisiones complejas o recomendaciones basándose en un amplio corpus de conocimiento, en muchas ocasiones difícil o imposible de revisar para los seres humanos. Los sistemas expertos se usan, por ejemplo, en el sector médico para la diagnosis y la planificación

de tratamientos. Durante mucho tiempo este fue el tipo de software de IA con más éxito.

Hoy en día la IA simbólica sigue siendo útil, pero han aparecido nuevos tipos de IA, que pueden (o no) combinarse con ella y que a diferencia de los sistemas expertos son capaces de aprender autónomamente a partir de datos. Esto se logra mediante un enfoque completamente distinto. El paradigma de investigación del *conexionismo* —que se desarrolló en los años 80 como una alternativa a lo que acabó llamándose la Inteligencia Artificial Anticuada (GOFAI, del inglés Good Old-Fashioned Artificial Intelligence)—, y la tecnología de las *redes neuronales* se basan en la idea de que, en lugar de representar funciones cognitivas superiores, necesitamos construir redes interconectadas basadas en unidades simples. Sus defensores afirman que es similar al modo en que funciona el cerebro humano: la cognición surge de interacciones entre unidades de procesado simples, llamadas «neuronas» (que, sin embargo, no son como las neuronas biológicas) y se utilizan muchas interconexiones neuronales. Este enfoque y esta tecnología se usan a menudo para el *aprendizaje automático* (véase el capítulo siguiente), también denominado *aprendizaje profundo (deep learning)* cuando las redes neuronales tienen varias capas de neuronas. Algunos sistemas son híbridos; AlphaGo de DeepMind, por ejemplo, lo es. El aprendizaje profundo ha permitido progresos en campos como el de la visión artificial o el del procesamiento de lenguajes naturales. El aprendizaje automático que emplea una red neuronal puede convertirse en una «caja negra» en el sentido de que, aunque los programadores conocen la arquitectura de la red, no está claro para otras personas qué es lo que ocurre en sus capas intermedias (entre el *input* y el *output)* y, por tanto, tampoco cómo se alcanza una decisión. Esta situación contrasta con la de los árboles de decisión, que son transparentes e interpretables y, por tanto, pueden ser revisados y evaluados por seres humanos.

Otro paradigma importante en la IA es el que usa enfoques corporeizados y situacionales, centrándose en las tareas motoras y en las interacciones más que en las llamadas tareas cognitivas superiores. Los robots construidos por los investigadores de IA, tales como el Rodney Brooks del MIT, no resuelven problemas usando representaciones simbólicas, sino interactuan-

do con su entorno circundante. Por ejemplo, Cog, el robot humanoide construido por Brooks y desarrollado en los años 90, fue diseñado para que aprendiese a interactuar con el mundo tal como lo hacen los niños. Además, no son pocos los que piensan que la mente solo puede surgir de la vida. Así, para crear una IA, necesitamos intentar crear vida artificial. Algunos ingenieros adoptan un enfoque menos metafísico y más práctico: toman la biología como modelo a partir del cual desarrollar las aplicaciones tecnológicas prácticas. También hay IAs evolutivas que tienen la capacidad de evolucionar. Algunos programas, empleando los llamados algoritmos genéticos, pueden incluso cambiarse a sí mismos.

Esta diversidad de enfoques y funciones de la IA implica también que hoy en la actualidad existan varios *subcampos:* aprendizaje automático, visión artificial, procesamiento de lenguajes naturales, sistemas expertos, computación evolutiva, *etc.* Actualmente el énfasis a menudo se pone en el aprendizaje automático, pero esta es solo una área de la IA, incluso si estas otras áreas están conectadas a menudo con el aprendizaje automático. Se ha alcanzado un gran progreso recientemente en la visión artificial, el procesamiento de lenguajes naturales y en el análisis de grandes cantidades de datos mediante el aprendizaje automático. Este último puede utilizarse, por ejemplo, para procesar lenguajes naturales basándose en el análisis de fuentes habladas y escritas, como textos extraídos de internet. Este tipo de trabajo es el que dio lugar a los agentes conversacionales actuales. Otro ejemplo es el reconocimiento facial basado en la visión artificial y en el aprendizaje profundo, que puede usarse, por ejemplo, para vigilancia.

Aplicaciones e impacto

La tecnología de IA se puede aplicar en distintos campos, desde la fabricación industrial, pasando por la agricultura y el transporte, hasta el sistema sanitario, las finanzas, el *marketing,* el sexo y el entretenimiento, la educación y las redes sociales. En las ventas al por menor y el *marketing,* se usan recomendadores o sistemas de recomendación para influir en las decisiones de compra y para ofrecer publicidad dirigida a un objetivo concreto. En las redes sociales, la IA es capaz de impulsar *bots:* cuentas de usuarios que pa-

recen ser de gente real pero que son, en realidad, software. Estos *bots* pueden publicar mensajes de contenido político o conversar con usuarios humanos. En atención médica, se emplea la IA para analizar datos de millones de pacientes. Los sistemas expertos también se siguen utilizando en dicha área. En finanzas, la IA se utiliza para analizar grandes conjuntos de datos para el análisis de mercado y los sistemas automáticos de *trading.* Los robots (de compañía) incluyen a menudo IA. Los pilotos automáticos y los coches autónomos usan IA. Los empresarios pueden usar IA para monitorizar a sus empleados. Los videojuegos tienen personajes que funcionan con IA. Las IAs pueden componer música o escribir artículos de periódico. También pueden imitar voces de personas e, incluso, falsificar discursos.

Dadas sus numerosas aplicaciones, es probable que la IA tenga un impacto generalizado, ahora y en un futuro próximo. Considérese la policía predictiva y el reconocimiento de habla, que crean nuevas posibilidades de seguridad y vigilancia, el transporte *peer-to-peer* y los coches autónomos que pueden transformar ciudades enteras, el *trading* algorítmico de alta frecuencia que ya afecta a los mercados financieros, o las aplicaciones diagnósticas en el sector médico que influyen en la toma de decisiones de los especialistas. No deberíamos olvidar que el de la ciencia es uno de los sectores en los que más impacto tiene la IA: mediante el análisis de grandes conjuntos de datos, la IA puede ayudar a los científicos a descubrir conexiones que, de otra manera, habrían pasado por alto. Esto es aplicable en ciencias naturales como la física, pero también en ciencias sociales y en humanidades. La IA ha afectado con seguridad al campo emergente de las humanidades digitales, por ejemplo, pues nos permite aprender cosas nuevas de los seres humanos y sus sociedades.

La IA tiene también un impacto en las relaciones sociales y una influencia amplia en la sociedad, la economía y el medio ambiente (Jansen *et al*. 2018). Es probable que condicione las interacciones humanas y tenga un impacto en la privacidad. Se dice que potencialmente incrementará el prejuicio y la discriminación. Se ha predicho que conducirá a la pérdida de empleo y, quizás, a la transformación de la economía al completo. Podría incrementar la brecha entre ricos y pobres y entre poderosos e indefensos, acelerando la injusticia y la desigualdad. Las aplicaciones militares pueden

cambiar la forma en que se libran las guerras gracias al desarrollo de armas letales automáticas. También tenemos que tener en cuenta el impacto medioambiental, que incluye el incremento del consumo de energía y de la contaminación. Más adelante examinaré con más detalle algunas de sus implicaciones éticas y sociales, centrándome en los problemas y en los riesgos de la IA. Pero es probable que la IA también nos traiga cosas positivas; por ejemplo, puede generar nuevas comunidades a través de las redes sociales, reducir tareas repetitivas y peligrosas al permitir que los robots las lleven a cabo, mejorar la cadena de suministros, reducir el consumo de agua, *etc.*

Pero no solo debemos preguntarnos por la naturaleza y el alcance de su impacto (positivo o negativo); es también importante preguntarse *a quién* afecta y de qué manera. Un impacto concreto puede resultar más positivo para unos que para otros. Hay muchas partes interesadas, desde trabajadores, pacientes y consumidores a gobiernos, inversores y empresas, y cada una de ellas puede verse afectada de forma distinta. Y estas diferencias en ganancias y vulnerabilidad ante los impactos de la IA surgen no solo dentro de los Estados, sino también entre países y regiones del mundo. ¿Beneficiará principalmente la IA a los países altamente avanzados y desarrollados? ¿Podría también beneficiar a gente menos educada y con menos ingresos, por ejemplo? ¿Quién tendrá acceso a la tecnología y será capaz de obtener sus beneficios? ¿Quiénes serán capaces de fortalecerse gracias a la IA? ¿A quién se excluirá de estos logros?

La IA no es la única tecnología digital que plantea estos problemas. Otras tecnologías de la información y de la comunicación digital también tienen un enorme impacto en nuestras vidas y sociedades. Como veremos, algunos problemas éticos de la IA no son exclusivos de esta. Por ejemplo, hay paralelismos con otras tecnologías de automatización. Considérense los robots industriales que, sin tener estatus de IA, generan desempleo. Y algunos de los problemas de la IA están relacionados con tecnologías asociadas a ella, como las redes sociales e internet, que, cuando se combinan con la IA, plantean nuevos desafíos. Por ejemplo, cuando las plataformas de redes sociales como Facebook la utilizan para saber más sobre sus usuarios, surge la preocupación por la privacidad.

¿Quién tendrá acceso a la tecnología y será capaz de obtener sus beneficios? ¿Quiénes serán capaces de fortalecerse gracias a la IA? ¿A quién se excluirá de estos logros?

Esta relación con otras tecnologías también significa que, a veces, la IA no es visible. Esto ocurre, en primer lugar, porque se ha convertido en una parte que ya está engranada en nuestra vida cotidiana. La IA se usa a menudo en aplicaciones nuevas y llamativas como AlphaGo. Pero no debemos olvidar que la IA ya impulsa plataformas de redes sociales, motores de búsqueda y otros medios y tecnologías que se han convertido en parte de nuestra experiencia de la vida cotidiana. La IA está en todas partes. La línea entre IA propiamente dicha y otras formas de tecnología puede ser borrosa, convirtiendo a la IA en invisible: si los sistemas de IA están integrados en las tecnologías, no solemos percatarnos su presencia. Y si sabemos que la IA está implicada, entonces es difícil decir si es la IA la que crea el problema o impacto, o si es la otra tecnología la que lo hace. En cierto sentido, no existe una «IA» en sí misma: la IA siempre se basa en otras tecnologías y está integrada en prácticas y procedimientos científicos y tecnológicos más amplios. Puesto que la IA también hace que surjan sus propios problemas éticos específicos, cualquier «ética de la IA» necesitará estar conectada con una ética más general de las tecnologías de la información y de la comunicación digitales, con una ética de la informática, *etc.*

No debemos olvidar que la IA ya impulsa plataformas de redes sociales, motores de búsqueda y otros medios y tecnologías que se han convertido en parte de nuestra experiencia de la vida cotidiana. La IA está en todas partes.

Otro sentido en el que podemos afirmar que la IA no es algo que exista por sí solo es que la tecnología es siempre social y humana: la IA no tiene que ver solamente con la tecnología, sino también de lo que los humanos hacen con ella, cómo la usan, cómo la perciben y la experimentan, y cómo la integran en entornos sociotécnicos más amplios. Esto es importante para la ética (que trata también las decisiones humanas) e implica que en ella debe incluirse una perspectiva histórica y sociocultural. El actual revuelo que la IA genera en los medios no es el primero que provoca una tecnología puntera. Antes de la IA, las palabras clave eran «robots» o «máquinas». Y otras tecnologías avanzadas como la nuclear, la nanotecnología, internet y la biotecnología también han producido muchos debates. Merece la pena tener esto en la cabeza cuando se discute sobre la ética de la IA, ya que quizás podamos aprender algo de estas controversias. El uso y el desarrollo de la tecnología se producen en un contexto social. Como sabe la gente que trabaja en evaluación de la tecnología, cuando esta es nueva tiende a ser muy controvertida, pero una vez que la tecnología se integra en la vida cotidiana, la exageración y la controversia se desinflan significativamente. Es probable que esto también suceda con la IA. Si bien es cierto que dicha predicción no es una buena razón para abandonar la tarea de evaluar los aspectos éticos y sociales de las consecuencias de la IA, nos ayuda a ver la IA en contexto y, en consecuencia, a comprenderla mejor.

Capítulo 6

Que no se nos olvide la ciencia de datos

Aprendizaje automático

Dado que muchas cuestiones éticas sobre la IA conciernen a tecnologías que están basadas entera o parcialmente en el aprendizaje automático y están relacionadas con la ciencia de datos, merece la pena acercarse a esta tecnología y a esta ciencia.

El *aprendizaje automático* se refiere al software que puede «aprender». El término es controvertido: algunos dicen que no es aprendizaje real porque no tiene una cognición real: solo los humanos pueden aprender. En cualquier caso, el aprendizaje automático moderno guarda «poca o ninguna similaridad con lo que plausiblemente podría estar ocurriendo en las cabezas humanas» (Boden 2016, pág. 46). El aprendizaje automático está basado en la estadística: es un proceso estadístico. Puede usarse para varias tareas, pero la tarea subyacente es a menudo el reconocimiento de patrones. Los algoritmos pueden identificar patrones o reglas observando conjuntos de datos y utilizar estos patrones o reglas para explicarlos y hacer predicciones.

Esto se consigue autónomamente en el sentido de que ocurre sin instrucciones ni reglas directas dadas por los programadores. En contraste con los sistemas expertos, que dependen de especialistas humanos en el campo en cuestión, quienes explican las reglas a los programadores que, a su vez, codifican estas reglas, el algoritmo de aprendizaje automático encuentra reglas o patrones que el programador no ha especificado. Solo se le da un objetivo o tarea. El software puede adaptar su comportamiento para cumplir mejor

con los requisitos de la tarea. Por ejemplo, el aprendizaje automático puede ayudar a distinguir correo no deseado de e-mails importantes buscando entre un gran número de mensajes y aprendiendo cuáles cuentan como correo no deseado. Otro ejemplo: para desarrollar un algoritmo que reconoce imágenes de gatos, el programador no da un conjunto de reglas al ordenador para definir lo que son los gatos, sino que hace que el algoritmo cree su propio modelo de imágenes de gatos. Se optimizará para alcanzar la mayor tasa de predicción en un conjunto de imágenes de gatos y no-gatos. Así, el algoritmo trata de aprender a reconocer las imágenes de gatos. Los humanos le dan retroalimentación, pero no le aportan instrucciones o reglas específicas.

Los científicos solían crear teorías para explicar datos y hacer predicciones; en el aprendizaje automático, los ordenadores crean sus propios modelos que encajan con los datos. El punto de partida son los datos, no las teorías. En este sentido, los datos dejan de ser «pasivos» y pasan a ser «activos»: son los «datos mismos los que definen lo que hacer a continuación» (Alpaydin 2016, 11). Los investigadores entrenan al algoritmo usando conjuntos de datos existentes (por ejemplo, viejos e-mails) y a continuación el algoritmo puede predecir resultados a partir de nuevos datos (por ejemplo, nuevos mensajes recibidos) (CDT 2018). Identificar patrones en grandes cantidades de información *(big data)* es lo que se llama a menudo «minería de datos» *(data mining),* haciendo una analogía con la extracción de minerales valiosos de la tierra. Sin embargo, el término es engañoso porque la meta es la extracción de patrones a partir de los datos, es decir, de su análisis; no la extracción de los datos mismos.

El aprendizaje automático puede estar *supervisado,* lo que significa que el algoritmo se centra en una variable particular que se designa como blanco para la predicción. Por ejemplo, si el objetivo es dividir a la gente en categorías (por ejemplo, riesgo de seguridad alto o bajo), las variables que predicen estas categorías ya son conocidas, y el algoritmo aprende entonces a predecir la categoría de pertenencia (riesgo de seguridad alto/riesgo de seguridad bajo). El programador entrena al sistema dándole ejemplos y contraejemplos, tales como imágenes de gente que suponen un alto riesgo de seguridad y de personas que no lo suponen. La meta es que el sistema aprenda a predecir quién pertenece a cada categoría, quién supone un riesgo

de seguridad alto y quién no, basándose en nuevos datos. Si se le otorgan al sistema los ejemplos suficientes, será capaz de generalizar a partir de estos ejemplos y sabrá categorizar nuevos datos, tales como una imagen nueva de un pasajero que está pasando un control de seguridad en un aeropuerto. Si está *no supervisado* quiere decir que este tipo de entrenamiento no se hace y que las categorías no son conocidas: los algoritmos crean sus propios grupos *(clusters).* Por ejemplo, la IA crea sus propias categorías de seguridad basándose en variables que ella misma selecciona; el programador no las facilita. La IA puede encontrar patrones que los especialistas en el campo (en este caso, personal de seguridad) aún no ha identificado. Es posible que sus categorías parezcan completamente arbitrarias para los humanos. Puede que no tengan sentido, pero cabe identificarlas estadísticamente. En otras ocasiones tienen sentido, y entonces este método nos permite adquirir nuevos conocimientos sobre las categorías del mundo real. El *aprendizaje por refuerzo,* finalmente, requiere que se indique si el *output* es bueno o malo. Su lógica es análoga a la de la recompensa y el castigo. No se le dicen al programa las acciones que debe llevar a cabo, sino que este «aprende» a través de un proceso iterativo en el que las acciones producen recompensas. Retomando el ejemplo de la seguridad: el sistema recibe retroalimentación de (datos suministrados por) el personal de seguridad para que «sepa» si ha hecho un buen trabajo en el caso de una predicción particular. Si una persona de la que se ha predicho que plantea un riesgo de seguridad bajo no causa ningún problema de seguridad, el sistema obtiene la retroalimentación de que este *output* era correcto y «aprende» de él. Nótese que siempre hay un porcentaje de error: el sistema nunca es cien por cien preciso. Nótese también que los términos técnicos «supervisado» y «no supervisado» tienen poco que ver con cómo de involucrados estén los humanos en el uso de la tecnología: aunque al algoritmo se le dé autonomía, en los tres casos los humanos estarán involucrados de alguna manera.

Esto también se aplica a todo lo que respecta al uso de datos en IA, incluido el llamado *big data.* El aprendizaje automático basado en el *big data* ha despertado mucho interés debido a la disponibilidad de grandes cantidades de datos y a un incremento en la potencia de los ordenadores. Algunos investigadores hablan de un «terremoto de datos» (Alpaydin 2016, pág. x).

Todos producimos datos derivados de nuestras actividades digitales, por ejemplo, cuando usamos las redes sociales o cuando compramos productos *online.* Estos datos son de interés para los actores comerciales, pero también para los gobiernos y los científicos. Nunca la recogida, el almacenaje y el procesamiento de datos habían sido tan sencillos para las organizaciones (Kelleher y Tierney 2018). Este hecho no se debe solamente al aprendizaje automático: el entorno digital más amplio y otras tecnologías digitales han desempeñado un importante papel. Las aplicaciones *online* y las redes sociales hacen más sencillo recolectar datos de los usuarios. Además, es más barato almacenar datos y los ordenadores se han vuelto más potentes. Todos estos factores han sido importantes para el desarrollo de la IA en general, pero también para la ciencia de datos.

Ciencia de datos

El aprendizaje automático está así vinculado a la *ciencia de datos.* Esta busca extraer patrones útiles y significativos de los conjuntos de datos, que actualmente son enormes. El aprendizaje automático es capaz de analizar estos grandes conjuntos. Tanto este como la ciencia de datos están basados en la estadística, que parte de observaciones particulares para alcanzar descripciones generales. Los especialistas están interesados en encontrar correlaciones en los datos. El modelado estadístico busca las relaciones matemáticas entre el *input* y el *output,* tarea para la cual el aprendizaje automático es de gran utilidad.

Pero la ciencia de datos supone más que su simple análisis mediante el aprendizaje automático. Los datos tienen que ser recogidos y preparados antes de analizarse, y después tienen que interpretarse los resultados de los análisis. La ciencia de datos incluye desafíos tales como el de obtener y limpiar los datos (por ejemplo, de las redes sociales y de la web), conseguir los datos suficientes, agrupar conjuntos de datos, reestructurarlos, seleccionar los relevantes y cuál es el tipo que vamos a utilizar. De este modo, los humanos desempeñan aún un papel importante en todas las fases y en relación con todos estos aspectos, incluyendo el encuadre del problema, la obtención de datos, la preparación de estos (el conjunto de datos con el que se

entrena al algoritmo y el conjunto al que se lo destinará), la creación o selección del algoritmo de aprendizaje, la interpretación de los resultados y la elección de una acción (Kelleher y Tierney 2018).

Todos producimos datos derivados de nuestras actividades digitales, por ejemplo, cuando usamos las redes sociales o cuando compramos productos *online*.

En cada fase de este proceso se presentan desafíos científicos concretos, y aunque el *software* puede ser sencillo de utilizar, el conocimiento especializado humano es necesario para lidiar con ellos. Normalmente es también necesaria la colaboración entre humanos, entre científicos de datos e ingenieros. Los errores siempre son posibles, y la elección humana, el conocimiento y la interpretación son cruciales. Los humanos son necesarios para interpretar de manera significativa y dirigir la tecnología hacia la búsqueda de diferentes factores y relaciones. Como subraya Boden (2016), la IA carece de nuestra comprensión de la relevancia. Se debería añadir que también carece de entendimiento, experiencia, sensibilidad y sabiduría. Este es un buen argumento en defensa de la idea de que en teoría y en principio deben estar involucrados los humanos. Pero también existe un argumento empírico a favor de la inclusión humana: en la práctica, los humanos *están* involucrados. Sin los programadores ni los científicos de datos, la tecnología, simplemente, no funciona. Además, la experiencia humana y la de la IA a menudo se combinan, por ejemplo, cuando un médico acepta una terapia contra el cáncer sugerida por una IA, pero también por su experiencia y su intuición de especialista. Si se deja fuera la intervención humana, las cosas pueden salir mal, carecer de sentido o, simplemente, resultar ridículas.

Tómese, por ejemplo, el conocido problema de la estadística, que también afecta a la utilización del aprendizaje automático de la IA: las correlaciones no son necesariamente relaciones causales. El libro de Tyler Vigen *Spurious Correlations* (2015) nos ofrece algunos ejemplos ilustrativos. En estadística, una correlación espúrea es aquella en la que las variables no están causalmente relacionadas, pero parecen estarlo, pues se deben a la presencia de un tercer factor invisible. Entre los ejemplos se incluye la correlación entre la ratio de divorcio en Maine y el consumo *per capita* de margarina, o la correlación entre el consumo *per capita* del queso mozzarella y los doctorados en ingeniería civil . Una IA puede encontrar estas correlaciones, pero los seres humanos son necesarios para decidir qué correlaciones merecen un estudio posterior a la hora de encontrar relaciones causales.

Además, ya en la fase de recolección de datos y en el diseño o creación de su conjunto se toman decisiones en tanto que se decide el modo por el que estos se extraen (Kelleher y Tierney 2018). La abstracción a partir de la realidad nunca es neutral, y la abstracción en sí misma no es la realidad, sino una representación. Esto quiere decir que podemos discutir cómo de buena y apropiada es una representación dado un propósito particular. Compárese esto con un mapa: el mapa en sí mismo no es el territorio, y los humanos tienen que tomar decisiones al diseñar el mapa con un propósito particular (por ejemplo, un mapa para la conducción de coches o un mapa topográfico para senderismo). En el aprendizaje automático, la abstracción mediante métodos estadísticos crea un modelo de la realidad: no es la realidad. También incluye elecciones: respecto al algoritmo mismo que provee la operación estadística que nos lleva de los datos al patrón/regla, hasta elecciones relativas al diseño del conjunto de datos con el que se entrena al algoritmo. Este aspecto de la elección del aprendizaje automático, de carácter humano, implica que podemos y debemos cuestionar críticamente las elecciones que se toman. Por ejemplo, ¿el conjunto de datos de entrenamiento es representativo de la población? ¿Hay sesgos en dichos datos? Como veremos en el capítulo siguiente, estas elecciones y preguntas nunca son meramente técnicas, sino que conllevan también un componente ético crucial.

Aplicaciones

El aprendizaje automático y la ciencia de datos tienen numerosas aplicaciones, algunas de las cuales ya he mencionado bajo el encabezado más general de la IA. Estas tecnologías pueden usarse para reconocer rostros (e incluso reconocer emociones basándose en el análisis de las caras), hacer sugerencias de búsquedas, conducir un coche, hacer predicciones de personalidad, predecir quién va a volver a delinquir o recomendar música. En ventas y en *marketing,* se usan para recomendar productos y servicios. Por ejemplo, cuando compramos algo en Amazon, el sitio recoge datos sobre nosotros y nos hace recomendaciones apoyándose en un modelo estadístico basado en los datos de todos los consumidores. Walmart ha probado la tec-

nología de reconocimiento facial para combatir los robos en sus tiendas; en el futuro podría usar la misma tecnología para determinar si los compradores están contentos o frustrados. Las tecnologías también tienen diversas aplicaciones en las finanzas. La empresa de crédito de referencia Experian trabaja con una IA de aprendizaje automático para analizar los datos sobre transacciones y las causas judiciales para determinar si se presta o no dinero al solicitante de una hipoteca. American Express usa aprendizaje automático para predecir transacciones fraudulentas. En el transporte, la IA y el *big data* se usan para crear coches autónomos. Por ejemplo, BMW usa un tipo de tecnología de reconocimiento facial para analizar datos que provienen de los sensores de los coches y de las cámaras. En atención médica, la IA de aprendizaje automático puede ayudar en el diagnóstico de cáncer (por ejemplo, analizando los escáneres radiológicos) o en la detección de enfermedades infecciosas. Por ejemplo, la IA DeepMind analizó un millón de imágenes de escáneres oculares y datos de pacientes con el fin de entrenarse para diagnosticar indicios de enfermedades oculares degenerativas. Watson, una IA de IBM, ha superado el juego de Jeopardy y se utiliza para dar recomendaciones para tratar el cáncer. Los complementos deportivos y los relacionados con la salud también ofrecen datos para las aplicaciones de aprendizaje automático. En el campo del periodismo, el aprendizaje automático sirve para escribir nuevas historias. Por ejemplo, en el Reino Unido la agencia de noticias Press Association tiene *bots* para las noticias locales. La IA también entra en el hogar y en la esfera privada, a menudo adoptando la forma de robots que recolectan datos y dispositivos de asistencia interactivos con procesamiento de lenguaje natural. Hello Barbie habla con los niños basándose en un sistema de procesamiento del lenguaje natural que analiza diálogos grabados. Todo lo que dicen los niños se graba, almacena y se analiza en los servidores de ToyTalk. A continuación, se envía una respuesta al dispositivo y Hello Barbie responde dependiendo de lo que haya «aprendido» acerca de su usuario. Facebook usa tecnologías de aprendizaje profundo y redes neuronales para estructurar y analizar datos de los casi dos mil millones de usuarios de la plataforma, que producen datos no estructurados. Esto ayuda a la empresa a ofrecer publicidad segmentada. Instagram analiza las imágenes de 800 millones de usuarios para vender publicidad a las empre-

sas. Mediante el uso de motores de recomendación que analizan los datos de los clientes, Netflix está pasando de ser un distribuidor a un creador de contenido: si eres capaz de adivinar lo que la gente quiere ver, puedes producirlo tú mismo y obtener beneficios. La ciencia de datos se ha usado incluso en cocina. Por ejemplo, basándose en el análisis de casi 10.000 recetas, el chef de IBM Watson crea sus propias recetas que sugieren nuevas combinaciones de ingredientes . La IA de aprendizaje automático también puede usarse en educación, contratación, derecho penal, seguridad (por ejemplo, policía predictiva), recuperación musical, trabajo de oficina, agricultura, armas militares, y mucho más.

La estadística no suele verse como un campo muy atractivo. Hoy, sin embargo, lo es por formar parte de la ciencia de datos y por la implicación de la IA y el *big data.* Es la nueva magia.

La estadística no suele verse como un campo muy atractivo. Hoy, sin embargo, lo es por formar parte de la ciencia de datos y por la implicación de la IA y el *big data.* Es la nueva magia. El tema preferido de los medios. Y un gran negocio. Algunos hablan de una nueva fiebre del oro: las expectativas son muchas. Además, este tipo de IA no es ciencia ficción o especulación; como muestran los ejemplos, la llamada IA estrecha o débil ya está aquí y lo invade todo. Por lo que se refiere a su impacto potencial, no hay nada estrecho o débil en ella. Es, por lo tanto, urgente analizar y discutir las muchas cuestiones éticas que surgen como consecuencia del aprendizaje automático y de otras tecnologías de IA y sus aplicaciones. De esto tratarán los siguientes capítulos.

Capítulo 7

La privacidad y otros sospechosos habituales

Muchos problemas éticos de la IA son conocidos por plantearse también en el área de la ética de la robótica y la automoción o, de forma más general, en el área de la ética de las tecnologías de la información y de la comunicación digitales. Pero esto, por sí mismo, no los convierte en menos importantes. Es más, gracias a ella en concreto y al modo en que se vincula a otras tecnologías, estas cuestiones toman una nueva dimensión y se vuelven incluso más urgentes.

Privacidad y protección de datos

Considérense, por ejemplo, la privacidad y la protección de datos. La IA, y en particular las aplicaciones de aprendizaje automático que operan con *big data,* a menudo implican la recogida y uso de información personal. La IA también puede utilizarse para vigilancia, tanto en la calle como en el lugar de trabajo y (gracias a *smartphones* y redes sociales) en cualquier otro sitio. A menudo la gente ni siquiera sabe que se están recogiendo esos datos, o que los datos que facilitaron en un contexto están siendo utilizados por terceras personas. Los conjuntos de *big data* a menudo implican la combinación de datos adquiridos por distintas organizaciones.

Un uso ético de la IA requiere que los datos sean recogidos, procesados y compartidos de una forma que respete la privacidad de los individuos y su derecho a saber lo que ocurre con ellos, al acceso de los mismos, a objetar su recogida o procesamiento, y a saber que se están recogiendo y procesando y (si procede) que están expuestos a la decisión tomada por una IA. Mu-

chas de estas cuestiones también surgen con otras tecnologías de la información y de la comunicación y, como veremos, la transparencia es un requisito igualmente importante (véase más adelante en este capítulo). También las cuestiones relativas a la protección de datos surgen en la investigación ética, por ejemplo, en la ética de la recolección de datos para la investigación en ciencias sociales.

Sin embargo, si se consideran los contextos en los que se usa la IA, estas cuestiones de privacidad y protección de datos se vuelven cada vez más problemáticas. Es relativamente fácil respetar estos valores y derechos cuando hacemos una encuesta como científicos sociales: se puede informar a los encuestados y pedir explícitamente su consentimiento, y lo que ocurrirá con los datos está bastante claro también. Pero el entorno en el que la IA y la ciencia de datos se están usando en la actualidad es muy distinto en general. Pensemos en las redes sociales: a pesar de la privacidad de la información y de las aplicaciones que piden a los usuarios su consentimiento, no queda claro para los usuarios lo que ocurre con los datos o incluso si se están recogiendo; y, si quieren utilizar la aplicación y disfrutar de sus beneficios, tienen que dar su consentimiento. A menudo, los usuarios tampoco saben siquiera si es una IA la que impulsa la aplicación. Y es frecuente que los datos dados en un contexto se muevan a otros dominios y se usen con propósitos diferentes (reutilización de datos), por ejemplo, cuando las empresas se los venden a otras o los mueven entre sus distintos departamentos sin que los usuarios lo sepan.

Manipulación, explotación y usuarios vulnerables

Este último fenómeno también apunta al riesgo de que los usuarios sean manipulados y explotados. La IA se usa para manipular lo que compramos, averiguar qué noticias seguimos, en qué opiniones confiamos, *etc.* Los investigadores en teoría crítica han señalado el contexto capitalista en el que tiene lugar el uso de las redes sociales. Por ejemplo, podría decirse que los usuarios de redes sociales realizan gratuitamente un «trabajo digital» (Fuchs 2014), pues producen datos para las empresas. Esta forma de explotación puede también involucrar a la IA. Como usuarios de las redes sociales, nos

arriesgamos a convertirnos en la mano de obra no pagada y explotada que produce datos para una IA que, a continuación, analiza nuestros datos; y en última instancia para las empresas que usan los datos, entre las que normalmente también se incluyen terceros. Esto recuerda a la advertencia de Herbert Marcuse en los años 60 de que incluso las sociedades llamadas «libres», «no totalitarias», tienen sus propias formas de dominación, en particular la explotación de los consumidores (Marcuse 1991). El peligro aquí consiste en que, incluso en las democracias actuales, la IA puede conducir a nuevas formas de manipulación, vigilancia y totalitarismo, no necesariamente bajo la apariencia de regímenes autoritarios, sino de manera subrepticia y altamente efectiva: cambiando la economía de manera que nos convierta en ganado de *smartphones* al que se ordeña para obtener datos. Pero la IA también puede usarse para manipular la política de forma directa, por ejemplo, analizando datos de las redes sociales para influir en campañas políticas (como en el famoso caso de Cambridge Analytica, una compañía que usó datos de usuarios de Facebook sin su consentimiento con fines políticos en las elecciones presidenciales de EEUU de 2016), o programando bots para que publiquen en redes sociales mensajes de carácter político basándose en los datos que reflejan las preferencias de los usuarios y así influir en una votación. Algunos también temen que la IA, al asumir tareas cognitivas propias de los seres humanos, infantilice a sus usuarios al «hacerlos menos capaces de pensar o de decidir por sí mismos lo que hacer» (Shanahan 2015, 170). Además, el riesgo de la explotación no se dirige solo al usuario: la IA depende del *hardware,* y su fabricación puede implicar la explotación de otras personas. La explotación puede estar también implícita en el entrenamiento de los algoritmos y en la producción de datos que se usan por y para la IA. Tal vez esta tecnología haga la vida de sus usuarios más fácil, pero este no es necesariamente el caso de aquellos que extraen los minerales, lidian con los desechos electrónicos y entrenan a la IA. Por ejemplo, el Alexa de Amazon Echo no solo genera un usuario que trabaja gratis, se convierte en fuente de datos, y se vende como un producto; un mundo de trabajo humano está también oculto detrás del escenario: mineros, transportistas marítimos, trabajadores del clic que etiquetan los conjuntos de datos,

todos al servicio de la importante acumulación de capital por parte de unos pocos (Schwab 2018).

La IA puede conducir a nuevas formas de manipulación, vigilancia y totalitarismo, no necesariamente bajo la apariencia de regímenes autoritarios, sino de manera subrepticia y altamente efectiva.

Algunos usuarios de la IA también son más vulnerables que otros. Las teorías de la privacidad y de la explotación suponen a menudo que el usuario es un ser humano autónomo, relativamente joven y sano, con capacidades mentales completas. Sin embargo, el mundo está también habitado por niños, ancianos, personas con discapacidad o con diversidades funcionales, *etc.* Estos usuarios vulnerables están más expuestos al riesgo. A menudo, puede violarse fácilmente su privacidad o son susceptibles de manipulación: la IA proporciona nuevas oportunidades para tales violaciones y manipulaciones. Considérense los niños pequeños que hablan con una muñeca que está conectada a un sistema tecnológico que incluye IA: lo más probable es que el niño no sepa que se está utilizando la IA, o que se están recogiendo datos, por no decir lo que se está haciendo con su información personal. Un *chatbot* o muñeca con IA no solo puede recoger de esta manera abundante información personal sobre el niño y sus padres: también puede manipular al niño por medio del lenguaje y una interfaz de voz. En la medida en que la IA se vuelve parte del «internet de los juguetes» (Druga y Williams 2017) y el internet de las (otras) cosas, esto se convierte en un problema ético y político. El fantasma del totalitarismo vuelve una vez más: no en historias de ciencia ficción distópicas o en pesadillas anticuadas de posguerra, sino por medio de tecnologías de consumo ya disponibles en el mercado.

«Fake news», la amenaza del totalitarismo y el impacto en las relaciones personales

La IA también puede utilizarse para generar discursos de odio e información falsa, o para crear *bots* que parezcan personas, pero que sean, en realidad, software de IA. Ya mencioné el *chatbot* Tay y el discurso falso de Obama. Este tipo de hechos podrían llevarnos a un mundo en el que ya no esté claro qué es cierto y qué es falso, donde se mezclen hechos y ficción. Tanto si es apropiado llamarlo «posverdad» como si no (McIntyre 2018), estas aplicaciones de la IA claramente contribuyen al problema. Por supuesto, la información falsa y la manipulación existen desde antes de la IA. Las pe-

lículas, por ejemplo, siempre han creado ilusiones, y los periódicos han difundido propaganda. Pero con la IA, combinada con las posibilidades y con el entorno de internet y de las redes sociales digitales, el problema parece haber empeorado considerablemente. Parece haber más oportunidades para la manipulación, lo cual pone en riesgo el pensamiento crítico. Todo esto nos recuerda, de nuevo, los peligros del totalitarismo, que se beneficia de la confusión en torno a la verdad y en el que se crean *fake news* con fines ideológicos.

Sin embargo, incluso en una utopía libertaria, las cosas puede que no parezcan tan prometedoras. La información falsa menoscaba la confianza y, por lo tanto, daña el tejido social. El uso generalizado de la tecnología puede conllevar una disminución del contacto entre personas o, al menos, del contacto significativo. Sherry Turkle (2011) ha denunciado lo siguiente con respecto a tecnologías tales como los ordenadores y los robots: terminamos esperando más de la tecnología, pero menos de los demás. Este argumento podría también aplicarse a la IA: la preocupación es que la IA, ya sea en forma de redes sociales o de «acompañantes» digitales, nos brinde la ilusión de la compañía, pero desestabilice las verdaderas relaciones con amigos, personas queridas y familias. A pesar de que esta preocupación ya existía antes de la IA y de que atañe siempre a los nuevos medios de comunicación (leer el periódico o ver la televisión, por ejemplo), el problema, con la IA, podría estar en que la tecnología es mucho *mejor* creando la ilusión de la compañía y, por tanto, incrementa el riesgo de la soledad o del deterioro de las relaciones sociales.

Protección y seguridad

Hay peligros más evidentes. También es necesario que las IAs no causen daño, especialmente cuando están integradas en sistemas de hardware que operan en el mundo físico. Considérense, por ejemplo, los robots industriales: se supone que no deben causar daño a los trabajadores; sin embargo, a veces ocurren accidentes en las fábricas. Los robots pueden matar, incluso si esto es relativamente raro. Sin embargo, con robots con IA, el problema de la seguridad se vuelve mucho más complejo: los robots deben ser capa-

ces de trabajar de forma más cercana con los humanos y de evitar causarles daños «inteligentemente». ¿Pero qué es lo que significa eso? ¿Deberían moverse más despacio cuando están cerca de un humano, lo que ralentizaría el proceso, o estaría bien que se movieran a gran velocidad para, así, realizar el trabajo eficientemente? Siempre existe el riesgo de que algo salga mal. ¿Las éticas de seguridad deberían reducirse a una cuestión de equilibrios? Los robots con IA en un entorno familiar o en espacios públicos también pueden generar problemas de seguridad. Por ejemplo, ¿debería siempre un robot evitar chocar contra humanos o es aceptable que a veces obstruya el paso a una persona para poder alcanzar su objetivo? Estas cuestiones no son meramente técnicas, tienen un componente ético: es un asunto de vidas humanas y valores como la libertad y la eficiencia. También hacen surgir problemas de responsabilidad (sobre esto hablaré más adelante).

Otro problema que ya existía antes de que la IA entrase en escena, pero que merece renovada atención, es el de la *seguridad.* En un mundo interconectado, cada dispositivo electrónico o software es susceptible de ser hackeado, invadido y manipulado por atacantes malintencionados. Todos conocemos, por ejemplo, los virus de ordenador, que nos los pueden estropear, pero si se equipan con IA, los dispositivos y el software pueden realizar otras operaciones complejas, y cuando adquieren capacidad de actuación y tienen incidencia en el mundo físico real, el problema de la seguridad se vuelve mucho más importante. Por ejemplo, si alguien hackea tu coche autónomo impulsado por IA, lo que tienes no es un simple «problema con el ordenador» o un «problema con el software»: tu propia vida está en juego. Y si alguien hackea el software de una infraestructura crítica (internet, agua, energía, etc.) o un dispositivo militar con capacidad letal, es probable que se desestabilice la sociedad entera y que muchas personas se vean afectadas. En aplicaciones militares, el uso de armas letales autónomas supone un riesgo de seguridad evidente; sobre todo, claro, para aquellos a los que se dirigen (normalmente no a habitantes de el mundo occidental), pero también para aquellos que las despliegan, pues siempre pueden ser hackeadas y volverse en su contra. Además, una carrera armamentística que implique estas armas podría conducir a una nueva guerra mundial. Y no es necesario mirar hacia un futuro muy lejano: si a día de hoy los drones (sin IA) ya pueden

inutilizar un gran aeropuerto de Londres, no es difícil imaginar cuán vulnerables son nuestras infraestructuras y cuán fácilmente el uso o el hackeo de la IA podrían causar trastornos masivos y damnificaciones. Nótese también que, en contraste con, por ejemplo, la tecnología nuclear, usar la tecnología de IA existente no requiere de un equipamiento caro o de un entrenamiento exhaustivo: las trabas para usar la IA con fines malintencionados son pocas.

En un mundo interconectado, cada dispositivo electrónico o software es susceptible de ser hackeado, invadido y manipulado por atacantes malintencionados.

Los problemas de seguridad mucho más mundanos con coches e infraestructuras como aeropuertos también nos recuerdan que, aunque hay gente más vulnerable que otra, *todos* somos vulnerables a tecnologías como la IA porque, al aumentar sus capacidades y nosotros delegar más tareas en ellas, nos volvemos más y más dependientes. Las cosas siempre pueden salir mal. Así pues, las nuevas vulnerabilidades tecnológicas nunca son solo tecnológicas: también se convierten en vulnerabilidades humanas, existenciales (Coeckelbergh 2013). Los problemas éticos discutidos aquí pueden, así, ser vistos como vulnerabilidades humanas: las vulnerabilidades tecnológicas transforman en última instancia nuestra existencia como humanos. En la medida en que nos volvemos dependientes de la IA, esta deja de ser una herramienta a nuestro servicio y se convierte en parte de nuestra naturaleza, así como de nuestra naturaleza ante el peligro en el mundo.

El incremento de las capacidades de la IA, especialmente cuando sustituyen a las humanas, también suscita otro problema ético incluso más urgente: la responsabilidad. Ese es el tema del siguiente capítulo.

Capítulo 8

Máquinas *arresponsables* y decisiones inexplicables

¿Cómo podemos y debemos atribuir responsabilidad moral?

Cuando la IA se usa para tomar decisiones y hacer cosas por nosotros, nos encontramos ante un problema compartido por todas las tecnologías de la automatización, pero que se vuelve aún más importante ahora que la IA nos permite delegar en las máquinas mucho más: la atribución de responsabilidad . Si a una IA se le otorga una mayor capacidad de actuación y asume tareas que solían llevar a cabo los humanos, ¿hemos de atribuirle responsabilidad moral? ¿Quién es responsable de los perjuicios y los beneficios que causa la tecnología cuando los humanos delegan capacidad de actuación y toma de decisiones en la IA? Reformulándolo en términos de riesgo: ¿quién es responsable cuando algo sale mal?

Cuando los seres humanos hacemos cosas y tomamos decisiones, normalmente vinculamos la capacidad de actuar con la *responsabilidad moral.* Eres responsable de lo que haces y de lo que decides.

Si a una IA se le otorga una mayor capacidad de actuación y asume tareas que solían llevar a cabo los humanos, ¿hemos de atribuirle responsabilidad moral?

Si lo que haces tiene efectos sobre el mundo y los demás, eres responsable de esas consecuencias. De acuerdo con Aristóteles, esta es la primera condición para la responsabilidad moral, la llamada condición de control: en la *Ética a Nicómaco* argumenta que la acción debe tener su origen en el agente. Esta visión tiene también un lado normativo: si tienes capacidad de actuar y puedes decidir, *deberías* ser responsable de tus actos. Lo que queremos evitar, hablando en términos morales, es que exista *alguien* con la capacidad de actuar y con el poder de hacerlo pero sin responsabilidades. Aristóteles añadió también otra condición para la responsabilidad moral: eres responsable si sabes qué es lo que estás haciendo. Esta es una condición *epistémica:* necesitas ser consciente de lo que estás haciendo y conocer las consecuencias que puede tener. Hemos de evitar la existencia de una entidad capaz de hacer, sin saber que las hace, cosas cuyos resultados pueden ser perjudiciales.

A continuación veremos cómo funcionan estas condiciones cuando delegamos decisiones y acciones en la IA. El primer problema es que la IA puede realizar acciones y tomar decisiones que tienen consecuencias éticas, pero no es consciente de lo que hace y no es capaz de pensamiento y moral y, por tanto, no puede hacerse moralmente responsable de lo que hace. Las máquinas pueden ser agentes, pero no agentes *morales,* ya que carecen de consciencia, libre albedrío, emociones, capacidad para formar intenciones y otras cualidades similares. Por ejemplo, desde la perspectiva aristotélica, solo los humanos pueden llevar a cabo acciones voluntarias y reflexionar sobre sus acciones. Si esto es cierto, la única solución consiste en hacer responsables a los humanos de lo que hace la máquina. Los humanos, entonces, delegarían la capacidad de actuar en la máquina, pero mantendrían la responsabilidad. En nuestro sistema legal ya hacemos esto: no hacemos responsables a los perros ni a los niños de sus acciones, sino que colocamos la responsabilidad legal en sus tutores. Y, en una organización, podemos delegar una tarea particular en un individuo, pero adscribir la responsabilidad a la persona a cargo del proyecto en general (a pesar de que, en este caso, aún habrá algo de responsabilidad por parte de aquel en quien se ha delegado) .

Entonces, ¿por qué no dejamos que las máquinas lleven a cabo las acciones y mantenemos la responsabilidad del lado del humano? Parece la solución más apropiada, ya que los algoritmos y las máquinas son *arresponsables.*

Sin embargo, esta solución se enfrenta a varios problemas en el caso de la IA. Primero, un sistema de IA puede tomar decisiones y llevar a cabo sus acciones muy rápidamente, por ejemplo, en transacciones de alta frecuencia o en un coche autónomo, lo que deja muy poco tiempo al humano como para tomar la decisión final o intervenir. ¿Cómo podrían ser responsables los humanos de acciones y decisiones de este tipo? Segundo, las IAs tienen historia. Cuando una hace algo en un contexto o aplicación particular, puede que ya no quede claro quién la creó, quién la utilizó primero, y cómo se debería distribuir la responsabilidad entre las diferentes partes involucradas. Por ejemplo, un algoritmo de IA creado en el contexto de un proyecto científico en una universidad puede encontrar su primera aplicación en el laboratorio de esa universidad, después en el sector sanitario y, por último, en un contexto militar. ¿Quién es responsable? Sería difícil rastrear a todos los seres humanos involucrados en la historia de esta IA particular y, ciertamente, en la historia causal que condujo hasta este resultado éticamente problemático. No siempre tenemos conocimiento de todas las personas que están involucradas en el momento en el que surge el problema de la responsabilidad. Un algoritmo de IA a menudo tiene una larga historia de la que numerosas personas han participado. Esto nos conduce a un problema típico relativo a la atribución de responsabilidad en actos llevados a cabo por tecnologías: normalmente hay demasiadas *manos* y, en mi opinión, demasiadas *cosas*.

Hay *demasiadas manos* en el sentido de que hay mucha gente involucrada en dicha acción tecnológica. En el caso de la IA, habríamos de empezar por el programador, pero también tenemos al usuario final y otros actores. Considérese, por ejemplo, el coche autónomo: están el programador, el usuario del coche, los propietarios de la empresa automovilística, los demás usuarios de la carretera, *etc.* En marzo de 2018 un coche autónomo de Uber causó un accidente en Arizona que tuvo como resultado la muerte de un peatón. ¿Quién es responsable de este trágico resultado? Podrían ser los que programaron el coche, los miembros de la empresa automovilística respon-

sables del desarrollo de producto, Uber, el usuario del coche, el peatón, el regulador (en este caso, el estado de Arizona), *etc.* No está claro quién es responsable. Puede que la responsabilidad no pueda y no se deba atribuir a una persona, sino a varias. Pero entonces no está claro cómo distribuir la responsabilidad. Algunas personas podrían cargar con una responsabilidad mayor.

También hay *muchas cosas,* en el sentido de que los sistemas tecnológicos consisten en numerosos elementos interconectados: normalmente hay involucrados diversos componentes del sistema. Está el algoritmo de IA, pero este algoritmo interactúa con sensores, usa todo tipo de datos, y todo tipo de *hardware* y *software.* Todas estas cosas tienen un recorrido y están vinculadas a la gente que las programó o produjo. Cuando algo sale mal, no siempre está claro si es «la IA» la que causó el problema o algún otro de los componentes del sistema (o, incluso, dónde acaba la IA y dónde empieza el resto de la tecnología), lo cual complica la tarea de adscribir y distribuir la responsabilidad. Además, hay que tener en cuenta el aprendizaje automático y la ciencia de datos: como hemos visto, no solo está el algoritmo, sino también un proceso que incluye varias etapas como la recopilación y tratamiento de datos, el entrenamiento del algoritmo, etc.; todos los cuales implican varios elementos técnicos y requieren decisiones humanas. De nuevo, existe una historia causal que implica a muchos humanos y a muchas partes; esto dificulta la cuestión de la atribución de responsabilidad.

Para intentar lidiar con estas cuestiones, se podría aprender de los sistemas legales o ver cómo funcionan los seguros; diré algo sobre nociones legales en el capítulo dedicado a las políticas. Pero detrás de esos sistemas legales y de seguros se esconden preguntas generales sobre la capacidad de actuar y la responsabilidad de la IA: ¿hasta qué punto queremos ser dependientes de la tecnología de la automatización?, ¿podemos responsabilizarnos de las acciones de la IA? y ¿cómo podemos atribuir y distribuir las responsabilidades? Por ejemplo, legalmente, la negligencia tiene que ver con el hecho de si una persona ha ejercido o no un deber de protección. ¿Pero en el caso de la IA, qué significa este deber, dado que es tan difícil predecir todas las potenciales consecuencias éticamente relevantes?

Esto nos lleva a la siguiente cuestión. Incluso si pudiéramos resolver el problema del control, también está la segunda condición relativa a la responsabilidad, que concierne al problema del conocimiento. Para ser responsable, uno necesita saber lo que está haciendo y provocando y, retrospectivamente, saber lo que uno ha hecho. Además, este problema tiene un aspecto relacional: en el caso de los humanos, esperamos que alguien pueda explicar lo que ha hecho o decidido. Responsabilidad significa entonces ser capaz de responder y dar explicaciones. Si algo sale mal, queremos una respuesta y una explicación. Por ejemplo, pedimos a un juez que justifique su decisión, o preguntamos a un delincuente por qué hizo lo que hizo. Todo esto genera importantes problemas en el caso de una IA, pues, en primer lugar, en principio la IA actual no «sabe» lo que está haciendo, en el sentido de que no tiene consciencia y que, por tanto, no es consciente de lo que está haciendo o provocando. Puede registrar y guardar lo que hace, pero no «sabe lo que está haciendo» de la misma manera que los humanos, quienes, por estar dotados de conciencia, saben lo que hacen y pueden (de nuevo, siguiendo a Aristóteles) discutir y reflexionar sobre sus acciones y sus consecuencias. Cuando no se cumplen estas condiciones en el caso de los humanos, en el caso de niños muy pequeños, por ejemplo, no los hacemos responsables. Normalmente tampoco hacemos responsables a los animales . En consecuencia, si la IA no cumple estas condiciones, no podemos hacerla responsable. La solución es, de nuevo, responsabilizar a los humanos de lo que hace la IA, asumiendo que *los humanos* saben lo que esta hace y lo que ellos hacen con ella, y (teniendo en cuenta el aspecto relacional) que *ellos* pueden responder por sus acciones y explicar lo que hace.

Sin embargo, si esta suposición es cierta, no se trata de algo tan sencillo como a primera vista puede parecer. Normalmente los programadores y los usuarios saben lo que quieren hacer con la IA, o de forma más precisa: saben lo que quieren que la IA haga por ellos. Saben la meta, el fin: por eso es por lo que delegan la tarea en la IA. Puede que también sepan cómo funciona la tecnología en general. Pero, como veremos, no *siempre* saben exactamente lo que está haciendo la IA (en cualquier momento del tiempo) y no *siempre* pueden explicar lo que hizo o cómo tomó su decisión.

Transparencia y explicabilidad

Aquí nos hallamos ante el problema de la *transparencia* y de la *explicabilidad.* Con algunos sistemas de IA, el modo por el cual toman las decisiones está claro. Si, por ejemplo, la IA usa un árbol de decisión, la forma en que llega a ella es transparente. Se ha programado de manera que determine su decisión, dado un *input* particular. Los seres humanos pueden, así, explicar cómo llegó la IA a esa decisión y por tanto es posible «pedir» a la IA que la «explique». Los humanos pueden hacerse responsables de la decisión o, más precisamente, de tomar una decisión basándose en las recomendaciones hechas por la IA. Sin embargo, en algunos otros sistemas de IA, sobre todo las IAs que usan aprendizaje automático y en particular el aprendizaje profundo de las redes neuronales, dicha explicación y dicho tipo de toma de decisiones ya no son posibles, pues el modo por el cual la IA decide deja de ser transparente y los humanos no pueden explicarla por completo. Saben cómo funciona el sistema, en general, pero no pueden explicar una decisión en particular. Pongamos por caso el aprendizaje profundo aplicado al ajedrez: el programador sabe cómo funciona la IA, pero la forma concreta en que la máquina llega a efectuar un determinado movimiento (esto es, lo que ocurre en las capas de la red neuronal) no es transparente y no se puede explicar. Esto constituye un problema en términos de responsabilidad, ya que los humanos que crean o usan la IA no pueden explicar una decisión concreta y, en consecuencia, no llegan a saber lo que está haciendo la IA, por lo que no pueden responder por sus acciones. Por un lado, los humanos saben lo que está haciendo la IA (por ejemplo, conocen el código de la IA y cómo funciona en general), pero por otro no (no pueden explicar una decisión concreta), de modo que a quien se vea afectado por las elecciones de la IA no se le podrá dar cuenta de las razones por las cuales esta tomó una u otra decisión. Así pues, y aunque todas las tecnologías de la automatización plantean problemas de responsabilidad, ciertos tipos de IA plantean uno especial: el llamado *problema de la caja negra.*

Además, ni siquiera la suposición de que en tales casos los humanos tenemos conocimiento sobre la IA en general y el código es siempre cierta. Probablemente los primeros programadores conozcan el código y cómo

funciona todo (o, al menos, conozcan la parte que ellos programaron), pero eso no significa que los siguientes programadores y usuarios que adaptan o utilizan los algoritmos para aplicaciones específicas comprendan perfectamente lo que está haciendo la IA. Por ejemplo, alguien que esté usando un algoritmo para temas de negocios puede no entender del todo la IA, o puede incluso no saber que la está utilizando, como en el caso de los usuarios de redes sociales. Y, por su parte, es posible que los primeros programadores desconozcan cuál es exactamente el futuro del algoritmo que desarrollan, o los diferentes campos de aplicación a los que este puede destinarse, por no hablar de los efectos *no deseados* de sus usos futuros. Así pues, independientemente del problema concreto del aprendizaje automático profundo, existe un problema de conocimiento con la IA en la medida en que mucha gente que utiliza la IA no sabe lo que está haciendo, pues desconoce lo que hace la IA, los efectos que tiene o incluso que la está usando. Este, también, es un problema de responsabilidad y, por tanto, un problema ético serio.

A veces estos problemas se plantean en términos de confianza: una falta de transparencia conduce a menos confianza en la tecnología y en los usuarios de dicha tecnología. Algunos investigadores se preguntan cómo incrementar entonces la confianza en la IA, e identifican la transparencia y la explicabilidad como factores que pueden incrementarla, así como, por ejemplo, el evitar sesgos (Winikoff 2018) y representaciones negativas de la IA, que la presenten como una especie de «Terminator» (Siau y Wang 2018). También, como veremos en el próximo capítulo, las políticas relativas a la IA conducen a menudo a generar confianza. Sin embargo, términos como IA «de confianza» son controvertidos: ¿deberíamos reservar la palabra «confiar» para relaciones entre humanos, o está bien usarla también para las máquinas? La investigadora de IA Joanna Bryson (2018) ha argumentado que la IA no es una cosa en la que se confíe, sino un conjunto de técnicas de desarrollo de *software:* cree que el término «confiar» debería estar reservado para la gente y sus instituciones sociales. Además, la cuestión de la transparencia y de la explicabilidad nos lleva a preguntarnos una vez más cuál es el tipo de sociedad que queremos. Aquí, el peligro no radica solamente en la manipulación y la dominación de las élites capitalistas o tecnocráticas, que crearían una sociedad enormemente dividida. El peligro adi-

cional y, quizás, más profundo que aquí se cierne es el de una sociedad altamente tecnológica en la que incluso las élites ignoran lo que están haciendo, y en las que nadie puede responder por lo que pasa.

Como veremos, los encargados de desarrollar políticas relativas a la IA proponen a veces una «IA explicable» y un «derecho a la explicación». Aun así, es cuestionable si es *posible* tener siempre IAs transparentes. Esto parece fácil de lograr en los sistemas clásicos. Pero si con las aplicaciones contemporáneas de aprendizaje automático resulta imposible, en principio, explicar cada paso del proceso de decisión y las decisiones relacionándolos con individuos concretos, tenemos un problema. ¿Se puede «abrir la caja negra»? Sería positivo, no solo para la ética, sino también para mejorar el sistema (esto es, el modelo) y aprender de él. Por ejemplo, si el sistema resultara explicable y la IA utilizase características que considerásemos inadecuadas, los seres humanos podríamos detectarlas y contribuir a eliminar las correlaciones espúreas. Y si una IA identificase nuevas estrategias para jugar a un juego y lo hiciese de manera transparente para los humanos, podríamos aprender de las máquinas a jugar mejor. Esto no solo es útil para los juegos, sino también en contextos como los del sistema sanitario, el derecho penal y la ciencia. Algunos investigadores, por tanto, intentan desarrollar técnicas para abrir la caja negra (Samek, Wiegand y Müller 2017). Pero si aún no es posible, o es posible solamente de forma limitada, ¿cómo debemos proceder? ¿Gira entonces el problema ético en torno al equilibrio entre rendimiento y explicabilidad? (Seseri 2018). Si un sistema es eficiente a costa de la transparencia, ¿deberíamos usarlo? ¿O deberíamos hacer lo posible por sortear el problema y dar con soluciones técnicas que permitan que hasta las IAs más avanzadas sean capaces de explicarse? ¿Es posible enseñar a las máquinas a hacer eso?

Además, incluso si la transparencia fuera deseable y posible, podría ser difícil de lograr en la práctica. Por ejemplo, las empresas privadas pueden no estar dispuestas a revelar sus algoritmos con tal de proteger sus intereses económicos. La legislación sobre propiedad intelectual que protege estos intereses también puede suponer un obstáculo. Y, como veremos en futuros capítulos, si la IA se encuentra en manos de poderosas corporaciones, se

plantean las cuestiones de quiénes son los que establecen las reglas de la IA y quiénes deberían ser.

Nótese, sin embargo, que éticamente hablando la transparencia y la explicabilidad no implican necesariamente —y, de hecho, en numerosas ocasiones no sucede así— revelar el código del *software.* Ante todo, lo que conlleva es explicar las decisiones; no explicar «cómo funciona» un sistema, sino cómo yo, en tanto que ser humano de quien se espera que rinda cuentas y actúe con responsabilidad, explico mi decisión. Cómo funciona y cómo llega la IA a su recomendación pueden ser parte de esa explicación. Por otro lado, la revelación del código, en sí misma, no facilita necesariamente información sobre el funcionamiento de la IA. Más bien, depende del bagaje educativo de la gente y de sus capacidades. Si las personas carecen de experiencia técnica relevante, se necesita de un tipo de explicación diferente. Esto no solo nos recuerda el problema de la educación, sino que nos conduce también a la cuestión de *qué tipo de explicación* se necesita y, en última instancia, qué es una explicación.

Así, el asunto de la transparencia y la explicabilidad plantea también preguntas filosóficas y científicas interesantes, como aquellas acerca de la naturaleza de la explicación (Weld y Bansal 2018). ¿Qué constituye una buena explicación? ¿Cuál es la diferencia entre explicaciones y razones?, ¿pueden las máquinas ofrecer unas y otras? Y, ¿cómo toman, en la práctica, decisiones los seres humanos? ¿Cómo explican sus decisiones? Se han dado investigaciones en torno a este asunto en psicología cognitiva y en ciencia cognitiva que podrían aplicarse a las IAs explicables. Por ejemplo, las personas generalmente no proporcionan cadenas causales completas. En su lugar, seleccionan explicaciones, y responden a lo que creen que son las creencias de aquel al que se lo explican: las explicaciones son sociales (Miller 2018). Y quizás esperemos también explicaciones por parte de las máquinas que difieran de las de los humanos, que solemos excusar ciertos errores amparándonos en nuestras emociones. Pero si hacemos tal cosa, ¿significa que consideramos superior el proceso de decisión de las máquinas a la toma de decisión humana? (Dignum *et al.* 2018), y si es así, ¿deberíamos hacerlo? Algunos investigadores hablan de razonar en vez de explicar. Winikoff (2018) incluso exige a las IAs y a otros sistemas autónomos

un «razonamiento basado en valores», que debería ser capaz de representar los valores y razones humanos por medio de valores humanos. ¿Pero puede «razonar» una máquina, y en qué sentido puede un sistema tecnológico «utilizar» o «representar» valores? ¿Tienen siquiera entendimiento? Y, como Boddington (2017) pregunta, ¿son los seres humanos verdaderamente capaces de establecer sin ambages sus valores fundamentales?

Tales problemas son interesantes para los filósofos, pero también tienen una relevancia ética directa y se dan en la realidad y en la práctica. Como señala Castelvecchi (2016): abrir la caja negra es un problema del mundo real. Por ejemplo, los bancos deberían explicar por qué rechazan conceder un préstamo; los jueces deberían explicar por qué envían a alguien (de vuelta) a prisión. Explicar decisiones no es solo una parte de lo que los humanos hacen naturalmente cuando se comunican (Goebel *et al.* 2018), es también una exigencia moral. La explicabilidad es una condición necesaria para el comportamiento y la toma de decisiones responsables. Parece razonable que todas las sociedades tengan a los seres humanos por individuos autónomos y sociales que intentan actuar y decidir responsablemente y que reclaman con acierto razones y explicaciones acerca de las decisiones que les afectan. Tanto si la IA puede proporcionar *directamente* estas razones y explicaciones como si no, los humanos deberíamos ser capaces de responder a la pregunta: «¿Por qué?». El desafío de los investigadores de la IA es asegurarse de que, si se utiliza una IA para un proceso de toma de decisiones, la tecnología esté construida de tal manera que los humanos sean capaces, en la medida de lo posible, de responder a esa pregunta.

Capítulo 9

El sesgo y el significado de la vida

Sesgo

Otro problema que es tanto ético como social, y también específico de la ciencia de datos basada en IA, como ciencia opuesta a otras tecnologías de la automatización, es la cuestión del sesgo. Cuando una IA toma (o, de forma más precisa, *recomienda)* decisiones, puede surgir el sesgo: las decisiones pueden ser injustas o poco equitativas para individuos o grupos particulares. A pesar de que puede surgir también en la IA clásica (pongamos, un sistema experto que utilice un árbol de decisión o una base de datos ya sesgados) el problema del sesgo está a menudo relacionado con las aplicaciones de aprendizaje automático. Y aunque los problemas de sesgo y discriminación siempre han estado presentes en la sociedad, la preocupación es que la IA pueda perpetuar estos problemas y acrecentar su impacto.

A menudo el sesgo no es intencionado: es habitual que los desarrolladores, usuarios y otros involucrados, como podría ser la dirección de una empresa, no prevean los efectos discriminatorios contra ciertos grupos o individuos.

Esto puede deberse a que no comprendan lo bastante bien el sistema de IA, a que no estén al tanto del problema del sesgo o, incluso, de su propio sesgo o, de forma más general, a que no piensen ni reflexionen lo suficiente sobre las consecuencias potenciales no intencionadas de la tecnología o se hayan apartado en exceso de determinados grupos de interés. Se trata de una cuestión problemática, ya que las decisiones que se toman con sesgos pueden tener consecuencias graves, por ejemplo, en términos de acceso a recursos y libertades (CDT 2018): los individuos podrían no conseguir un trabajo, no recibir un crédito, acabar en la cárcel o incluso padecer violen-

cia. Y no solo pueden sufrir los individuos; comunidades enteras son susceptibles de verse afectadas por decisiones sesgadas, por ejemplo, cuando los habitantes de cierta zona de una ciudad o todas las personas con determinadas raíces étnicas son catalogados por la IA como representantes de un alto riesgo de seguridad.

Aunque los problemas de sesgo y discriminación siempre han estado presentes en la sociedad, la preocupación es que la IA pueda perpetuar estos problemas y acrecentar su impacto.

Considérese de nuevo el ejemplo del algoritmo COMPAS, mencionado en el primer capítulo, que predice si ciertos acusados están más dispuestos que otros a volver a delinquir y que fue utilizado por jueces en Florida para decidir en sentencia judicial, por ejemplo, sobre cuándo se le otorgaría la libertad vigilada a una persona. De acuerdo con un estudio realizado por la redacción online de ProPublica, los falsos positivos del algoritmo (acusados de los que se predijo que iban a volver a delinquir pero que no lo hicieron) eran en una gran proporción de raza negra, y los falsos negativos (acusados de los que se predijo que no iban a volver a delinquir pero que sí lo hicieron) lo eran de raza blanca (Fry 2018). Los críticos argumentaron que había un prejuicio contra los acusados negros. Otro ejemplo es PredPol, la herramienta de policía predictiva que se ha empleado en los Estados Unidos para predecir la probabilidad de que se cometan delitos en zonas concretas de las ciudades y recomendar la distribución de los efectivos policiales (por ejemplo, en qué lugares deberían patrullar los oficiales de policía). La preocupación reside en este caso en que el sistema parece estar sesgado en contra de barrios pobres y personas racializadas, o en que una vigilancia policial desproporcionada demolería la confianza entre la gente en esas áreas, convirtiendo la predicción en una profecía autocumplida (Kelleher y Tierney 2018). Pero el sesgo no afecta solamente a la justicia penal o a la policía: también puede implicar, por ejemplo, que los usuarios de servicios de internet estén discriminados si la IA los categoriza desfavorablemente.

El sesgo puede surgir de diversas maneras en todas las fases de diseño, prueba y aplicación. Centrémonos en el diseño: en este caso puede surgir en la selección del conjunto de datos de entrenamiento; en el propio conjunto de datos de entrenamiento, por no ser representativo o por estar incompleto; en el algoritmo; en el conjunto de datos que se le da al algoritmo una vez que se lo ha entrenado; en decisiones basadas en correlaciones espúreas (véase el capítulo anterior); en el grupo que crea el algoritmo; o en la sociedad en general. Pongamos por caso un conjunto de datos no representativo de la población (por ejemplo, puede estar basado en hombres americanos blancos) que aun así se utilice para hacer predicciones sobre la población al

completo (hombres y mujeres de diferentes orígenes étnicos). Además, el sesgo también puede afectar a países. Muchas redes neuronales para el reconocimiento facial se entrenan con conjuntos de datos extraídos de ImageNet, que contienen una gran cantidad de datos de los Estados Unidos, mientras que países como China y la India, que representan a una parte mucho mayor de la población mundial, contribuyen solamente con una pequeña fracción (Zou y Schiebinger 2018), lo cual puede llevarnos a incorporar un sesgo cultural al conjunto de datos. De forma más general, los conjuntos de datos pueden estar incompletos o ser de mala calidad, lo que puede conducir al sesgo. La predicción puede también estar basada en una información insuficiente, por ejemplo, en el caso de la predicción de asesinatos: no hay tantos asesinos, por lo que la generalización es problemática. Otro ejemplo: algunos investigadores se preocupan por la falta de diversidad entre los desarrolladores de IA y los equipos de ciencia de datos: la mayoría de los informáticos e ingenieros son hombres blancos de países occidentales de entre 20 y 40 años, y su experiencia personal, opiniones y, por supuesto, prejuicios, pueden influir en su trabajo, afectando potencialmente de manera negativa a las personas que no encajan con esta descripción, como mujeres, discapacitados, gente mayor, gente de color y gente de países en vías de desarrollo.

Los datos también pueden estar sesgados en contra de grupos concretos porque el sesgo esté integrado en la práctica específica o en la sociedad en general. Considérense las afirmaciones de que la medicina usa datos principalmente de pacientes masculinos y, por tanto, están sesgados, o sesgados en perjuicio de personas racializadas, que son dominantes en la sociedad en general. Si un algoritmo se alimenta de estos datos, el resultado de sus actividades también lo estará. Si hay sesgo *dentro,* habrá sesgo *fuera,* como ya se advirtió en un editorial de *Nature* en 2016. También se ha demostrado que el aprendizaje automático puede adquirir sesgos si se alimenta de datos provenientes de internet, dado que dichos datos presentados en un lenguaje corriente reflejan la cultura humana cotidiana, incluyendo sus prejuicios (Caliskan, Bryson y Narayanan 2017). El lenguaje, por ejemplo, contiene sesgos de género. La preocupación estriba en que la IA pueda perpetuarlos, desfavoreciendo más aún a grupos históricamente marginados. El sesgo

puede surgir también si hay una correlación, pero no una causalidad. Tomando de nuevo un ejemplo de la justicia penal: un algoritmo puede inferir que, si uno de los padres del acusado fue a prisión, es más probable que este termine en la cárcel. Incluso si dicha correlación se diese, e incluso si la inferencia fuera predictiva, parece injusto que el acusado reciba una sentencia más dura, pues no existe una relación causal (House of Commons 2018). Finalmente, el sesgo también puede aparecer porque los humanos que toman las decisiones confían más en la precisión de las recomendaciones de los algoritmos de lo que debieran (CDT 2018) y menosprecian otra información o aplican su propio juicio menos de lo que convendría. Por ejemplo, un juez puede basarse enteramente en el algoritmo y no tener en cuenta otros elementos. Como ocurre siempre con la IA y con otras tecnologías de la automatización, las decisiones y la interpretación humana juegan un papel muy relevante, y siempre existe el riesgo de confiar demasiado en la tecnología.

Sin embargo, no está claro si el sesgo se puede evitar de alguna forma; ni siquiera si debería evitarse ni, en caso de que se pudiese, a qué coste habría que hacerlo. Por ejemplo, si modificar el algoritmo de aprendizaje automático para reducir el riesgo de sesgo hace que sus predicciones sean menos precisas, ¿deberíamos hacerlo? Puede que quepa encontrar un equilibrio entre la efectividad del algoritmo y contrarrestar el sesgo. También existe el problema de que, si ciertas características como la raza se ignorasen o eliminasen, los sistemas de aprendizaje automático podrían identificarlas igualmente a partir de variables representativas de dichas características, y eso también puede conducir al sesgo. Por ejemplo, en el caso de la raza podría ocurrir que otras variables que están correlacionadas con la raza, tales como el código postal, fueran seleccionadas por el algoritmo. ¿Es posible, pues, un algoritmo perfectamente imparcial? No hay un consenso entre los filósofos ni en la sociedad sobre lo que son la justicia y la equidad perfectas. Además, como señalé en el capítulo anterior, los conjuntos de datos utilizados por los algoritmos son abstracciones de la realidad y el resultado de elecciones humanas y, por tanto, nunca son neutrales (Kelleher y Tierney 2018). El sesgo impregna nuestro mundo y nuestras sociedades; por ello, a pesar de que podemos y debemos hacer todo lo posible por minimizar el

sesgo, los modelos de la IA nunca estarán completamente libres de él (Digital Europe 2018).

Además, sin duda los algoritmos usados para la toma de decisiones están *siempre* sesgados en el sentido de que son discriminatorios: deben discernir entre varias posibilidades. Por ejemplo, en un proceso de selección la revisión de los CV debe estar sesgada y ser discriminatoria respecto de aquellas características del candidato que mejor encaja en el puesto. Las cuestión ética y política es si una discriminación particular es injusta o desigual. Pero, de nuevo: las opiniones sobre justicia y equidad difieren. Esto hace que la cuestión del sesgo no sea solo técnica, sino que también esté relacionada con las discusiones políticas sobre justicia y equidad. Por ejemplo, es controvertido que una acción de discriminación positiva o afirmativa, que intenta neutralizar un sesgo creando otro sesgo positivo para con los individuos o grupos en desventaja, sea justa. ¿Debería ser la justicia ciega e imparcial (y, por tanto, deberían ser los algoritmos ciegos ante la raza, por ejemplo)?, ¿o debería tratar de compensar a quienes parten con una desventaja, procurando equilibrar (correctivamente) la parcialidad y la discriminación? Y, ¿debería la política en un contexto democrático priorizar la protección de los intereses de la mayoría o centrarse en fomentar los intereses de una minoría, ya sea una minoría histórica o una desvaforecida puntualmente?

Esto nos lleva al asunto de las soluciones. Incluso si estamos de acuerdo en que existe el sesgo, existen varias formas de lidiar con él. Entre estas se incluyen soluciones tecnológicas, pero también medidas sociales, políticas y educativas. La decisición de tomar unas u otras resulta controvertida, y depende, de nuevo, de nuestra noción de justicia y equidad. Por ejemplo, la cuestión acerca de la acción afirmativa conduce a una cuestión más amplia sobre si deberíamos aceptar el mundo como es o moldear activamente el futuro de manera que se evite la perpetuación de las injusticias del pasado. Algunos argumentan que se podría utilizar un conjunto de datos que reflejara el mundo real. Los datos pueden representar prejuicios de la sociedad y el algoritmo modelar sesgos preexistentes, pero este no es un problema del que deban preocuparse los desarrolladores. Otros argumentan que tales conjuntos de datos existen solo a causa de siglos de sesgos, que dichos sesgos y

discriminaciones son injustos y afectan desigualmente a las personas, y que, por tanto, se debería cambiar ese conjunto de datos, o el algoritmo, para poder fomentar la acción positiva. Por ejemplo, en respuesta a los resultados del algoritmo de búsqueda de Google que parece presentar un sesgo en favor de los hombres cuando se busca *«math professors»* (es decir, que devuelve muchas más imágenes de profesores que de profesoras), se podría argüir que estos, simplemente, reflejan el mundo tal y como es (y que dicho reflejo, que evidencia que históricamente ha habido más profesores varones que profesoras, es exactamente lo que debería proporcionar el algoritmo); o podríamos hacer que el algoritmo priorizara imágenes de profesoras de matemáticas para así cambiar la percepción y, quizás, cambiar el mundo (Fry 2018). Se podría también intentar formar equipos de desarrollo más diversos en términos de origen étnico, opinión y experiencia, y que representen mejor a los grupos potencialmente afectados por el algoritmo (House of Commons 2018).

¿Debería ser la justicia ciega e imparcial o debería tratar de compensar a quienes parten con una desventaja?

El punto de vista del espejo no funciona si los datos de entrenamiento no reflejan el mundo tal y como es, y tampoco si contiene datos viejos que no representen la situación actual. Las decisiones basadas en estos datos ayudarían, en ese caso, a perpetuar nuestro pasado discriminatorio en vez de preparar el futuro. Además, otra objeción contra el punto de vista del espejo es que, incluso si un modelo refleja el mundo tal y como es, podría llevar a la discriminación y al perjuicio con respecto a individuos y grupos específicos. Por ejemplo, siguiendo criterios generados por una IA, las entidades de crédito pueden denegar préstamos a los solicitantes basándose en su lugar de residencia, e igualmente ciertos comercios online pueden cobrar más a unos consumidores que a otros. Los perfiles pueden también seguir a los individuos a través de dominios (Kelleher y Tierney 2018). Y una simple función de autocompletado puede ligar falsamente tu nombre a un delito (con consecuencias terribles), incluso si la IA de búsqueda que se encuentra detrás refleja correctamente el mundo porque, supongamos, un gran número de personas busca por error tu nombre en lugar del del delincuente. Otro ejemplo, quizás menos obvio, de sesgo: un sistema de recuperación de música utilizado por servicios como Spotify, que hace recomendaciones a partir del comportamiento real de los usuarios (sigue la pista de los *clicks),* puede discriminar a la música y a los músicos menos populares. Incluso si reflejase el mundo tal y como es, conduciría a una situación en la que ciertos músicos no podrían vivir de su música y ciertas comunidades no se sentirían reconocidas ni respetadas.

Estos son claros casos de discriminación problemática, sin embargo, siempre deberíamos preguntarnos: ¿la discriminación en un caso particular es justa o no? Y si la consideramos injusta, ¿qué se puede hacer al respecto?, ¿y quién? Por ejemplo, ¿qué pueden hacer los informáticos? ¿Deberían generar los conjuntos de datos de entrenamiento más variados, quizás incluso creando datos y conjuntos «idealizados» como sugirió Eric Horvitz (Microsoft) (Surur 2017)? ¿O deberían reflejar el mundo? ¿Deberían propiciar los desarrolladores discriminaciones positivas o crear algoritmos «ciegos»? Cómo lidiar con el sesgo en la IA no es solo un asunto técnico: es una cues-

tión política y filosófica. Las claves están en qué tipo de sociedad y mundo queremos, si debemos intentar cambiarlos y, si es así, qué tipos de cambios son aceptables y equitativos. Se trata de un asunto que afecta tanto a seres huamnos como a máquinas: ¿creemos que la toma de decisiones *humana* es justa e imparcial y, si no, cuál es el papel de la IA? Quizás la IA nos pudiera enseñar cosas sobre los humanos y las sociedades humanas al descubrirnos nuestros sesgos. Además, discutir sobre ética de la IA puede revelar desequilibrios de poder sociales e institucionales.

En consecuencia, los debates sobre ética en IA tocan de lleno delicados problemas políticos y sociales que, a su vez, están ligados a problemas filosóficos de índole normativa, como los relativos a la justicia y la equidad, así como a problemas filosóficos y científicos relacionados con los seres humanos y sus sociedades. Uno de estos problemas es el del futuro del trabajo.

El futuro del trabajo y el significado de la vida

Se ha predicho que la automatización controlada por IA transformará radicalmente nuestras economías y sociedades, y ello nos llevará a plantearnos preguntas no solo sobre el futuro y el significado del trabajo, sino también sobre el futuro y el significado de la vida humana.

En primer lugar, está la preocupación por que la IA destruya puestos de trabajo, causando, quizás, un desempleo masivo. También está la cuestión de qué tipo de trabajo será aquel del que se apodere la IA: ¿solo trabajos propios de obreros, o también otros? Un famoso informe de Benedikt Frey y Michael Osborne (2013) predice que el 47 % de todos los trabajos de Estados Unidos podrían llegar a automatizarse. Otros informes proponen cifras menos espectaculares, pero la mayoría predicen que la pérdida de empleo será significativa. Muchos autores están de acuerdo en que la economía se ha transformado enormemente y continuará haciéndolo (Brynjolfsson y McAffee 2014), afectando al empleo sustancialmente, ahora y en el futuro. Además, se predice que la pérdida de empleo debida a la IA afectará a todo tipo de trabajadores, no solo obreros, pues la IA es cada vez más capaz de realizar tareas cognitivas complejas. Si eso fuera cierto, ¿cómo preparar a las nuevas generaciones para el futuro? ¿Qué deberían aprender?

¿Qué deberían hacer? ¿Y qué pasa si la IA beneficia más a unas personas que a otras?

Con esta última pregunta, tocamos de nuevo la cuestión de la justicia y la equidad, que ha tenido ocupados a los pensadores de la filosofía política durante mucho tiempo. Si la IA fuera a crear una brecha mayor entre ricos y pobres, por ejemplo, ¿sería esto justo? Y si no, ¿qué podemos hacer? También se podría plantear el problema en términos de desigualdad (¿incrementará la IA la desigualdad en las sociedades y en el mundo?) o en términos de vulnerabilidad: ¿disfrutará el empleado, rico y educado en países tecnológicamente avanzados, de los beneficios de la IA, mientras que el desempleado, pobre y menos educado en países en vías de desarrollo será mucho más vulnerable a sus impactos negativos? (Jansen *et al.* 2018). Y por retomar una preocupación ética y política más reciente: ¿qué pasa con la justicia medioambiental? ¿Cuál es el impacto de la IA en el medio ambiente y en nuestra relación con el entorno? ¿Qué significa «IA sostenible»? También está la cuestión de si la ética de la IA y la política deberían dedicarse en exclusiva a los valores y los intereses humanos (véase capítulo 12).

Otra pregunta con tintes existencialistas se refiere al significado del trabajo y de las vidas humanas. La preocupación por la destrucción de empleo presupone que el trabajo es el único valor y la única fuente de ingresos y de sentido. Sin embargo, si el trabajo es lo único que vale, ¿no deberíamos fomentar las enfermedades mentales, fumar más y engordar, ya que todos esos problemas generan puestos de trabajo? . Lo que queremos no es eso. Claramente, creemos que hay otros valores más importantes que la creación de empleo en sí misma. ¿Y por qué depender de los empleos para generar ingresos y sentidos? Podríamos organizar nuestras sociedades y economías de una forma diferente. Podríamos disociar el trabajo y los ingresos, o más bien lo que consideramos «trabajo» e ingresos. Hoy día mucha gente realiza trabajos no remunerados, por ejemplo, en el hogar y en el cuidado de los niños y de los mayores. ¿Por qué esto no se considera «trabajo»? ¿Por qué resulta menos significativo en términos de identidad hacer este tipo de trabajo? ¿Y por qué no hacemos de él una fuente de ingresos? Además, hay gente que piensa que la automatización nos podría permitir a algo mejor que lo que hoy entendemos por ocio. Quizá pudiéramos dedicarnos a cosas más

placenteras y creativas, no necesariamente en la forma de un trabajo. Es posible, en otras palabras, cuestionar la idea de que una vida plena de sentido pasa necesariamente por entregarse a un trabajo asalariado y pre-estructurado por terceros o por el marco del llamado trabajador autónomo. Quizás pudiéramos aplicar medidas tales como una «renta básica» para poder permitir a todo el mundo que haga lo que considere que le aporta un sentido. Así, en respuesta al problema del futuro del trabajo, podemos pensar en qué es lo que da sentido al trabajo, qué tipo de trabajo le estaría permitido hacer a los seres humanos, y cómo podemos reorganizar nuestras sociedades y economías de tal forma que los ingresos no estén limitados al trabajo y al empleo.

Se ha predicho que la automatización controlada por IA transformará radicalmente nuestras economías y sociedades, y ello nos llevará a plantearnos preguntas no solo sobre el futuro y el significado del trabajo, sino también sobre el futuro y el significado de la vida humana.

Dicho esto, hasta ahora las ideas utópicas sobre sociedades del ocio y otros paraísos postindustriales no se han llevado a cabo. Ya hemos tenido varias olas de automatización desde el siglo XIX hasta ahora, pero ¿hasta qué punto nos han liberado y emancipado las máquinas? Quizás se han hecho cargo de algunos trabajos sucios y peligrosos, pero también se han utilizado para la explotación y no han cambiado radicalmente la estructura jerárquica de la sociedad. Algunos se han beneficiado enormemente de la automatización, mientras que otros no. Quizás la fantasía de no trabajar sea un lujo reservado a quienes están del lado ganador. Por otra parte, ¿nos han liberado las máquinas hasta el punto de tener vidas con más sentido? ¿O amenazan la posibilidad misma de tales vidas? Esta es una discusión antigua y no hay respuestas fáciles a estas preguntas, pero las preocupaciones surgidas son buenos motivos para, al menos, ser escépticos sobre el nuevo mundo feliz que dibujan los profetas de la IA.

Además, quizás el trabajo no sea necesariamente un castigo que haya que evitar, o una explotación a la que resistirse: desde una perspectiva diferente se puede afirmar que el trabajo tiene valor, que otorga al trabajador un propósito y un sentido, y que aporta beneficios diversos, tales como vínculos sociales, la pertenencia a algo más grande, salud y oportunidades de ejercer una responsabilidad (Boddington 2016). Si este es el caso, entonces quizás podamos *reservar* el trabajo para los humanos; o, al menos, algún tipo de trabajo que aporte un sentido y que ofrezca oportunidades para que esos bienes se materialicen. O, al menos, algunas *tareas.* La IA no tiene por qué hacerse cargo de todo el trabajo, pero podría hacerlo de algunas tareas que no nos parezcan significativas. Podemos colaborar con las IAs. Por ejemplo, podríamos escoger no delegar trabajos creativos en las IAs (algo que propone Bostrom) o colaborar con las IAs para hacer cosas creativas. La preocupación aquí puede ser que, si las máquinas se hacen cargo de todo lo que hacemos ahora en la vida, no nos quedaría nada que hacer y nos hallaríamos ante una vida sin sentido. Sin embargo, lo dicho es totalmente incierto: teniendo en mente el escepticismo sobre lo que la IA puede hacer (véase el capítulo 3) y el hecho de que muchas de las actividades que no

consideramos «trabajo» nos aportan, sin embargo, un sentido, probablemente nos queden muchas cosas a las que dedicarnos. La cuestión no es entonces lo que los humanos harán cuando *todos* los trabajos los hagan las máquinas, sino qué tareas queremos o deberíamos reservar para los humanos y qué tipo de papel puede tener la IA, si es que tiene alguno, para ayudarnos de maneras ética y socialmente aceptables.

Para concluir, la ética de la IA nos hace pensar en cuál sería una sociedad buena y justa, qué es una vida humana con sentido y qué papel tiene y puede tener la tecnología en relación con estas cuestiones. La filosofía, incluyendo la filosofía antigua, puede perfectamente ser una fuente de inspiración para la reflexión sobre las tecnologías actuales y sus problemas éticos y sociales potenciales y *de facto.* Si la IA hace surgir de nuevo esas viejas preguntas sobre la vida buena y plena de sentido, tenemos recursos en diferentes tradiciones filosóficas y religiosas que pueden ayudarnos a abordarlas. Por ejemplo, como ha argumentado Shannon Vallor (2016), la tradición de la ética de la virtud desarrollada por Aristóteles, Confucio y otros pensadores antiguos puede ayudarnos todavía a reflexionar sobre el florecimiento humano y lo que debería suponer en una era tecnológica. En otras palabras, podría ser que ya tengamos las respuestas a esas preguntas, pero necesitemos reflexionar sobre lo que significa tener una buena vida en el contexto de las tecnologías actuales, incluyendo la IA.

Sin embargo, la idea de desarrollar «una ética de la IA para una buena vida» y una ética de la IA para el mundo real en general se enfrenta a numerosos problemas. El primero es el de la *velocidad.* El modelo de ética de la virtud que ha heredado de Aristóteles la filosofía Occidental presupone una sociedad que se va transformando lentamente, en la que la tecnología no cambia tan rápido y en la que la gente tiene tiempo para dedicarlo al conocimiento práctico; no está claro cómo podríamos utilizarla para hacer frente a una sociedad que cambia tan rápidamente (Boddington 2016) y al veloz desarrollo de tecnologías tales como la IA. ¿Tenemos todavía el tiempo que requiere responder, desarrollar y transmitir un conocimiento práctico relacionado con el uso de tecnologías como la IA? ¿Llega la ética demasiado tarde? Cuando Minerva, la lechuza de la filosofía, abra finalmente sus alas, puede que el mundo ya haya cambiado tanto que no se lo reconozca. ¿Cuál

es y cuál debería ser el papel de una ética así en el contexto de los desarrollos del mundo real?

En segundo lugar, dada la diversidad y pluralidad de perspectivas sobre este asunto dentro de las sociedades y las diferencias culturales entre sociedades, las cuestiones sobre una vida buena y con sentido coexistiendo con la tecnología pueden recibir diferentes respuestas en distintos sitios y contextos, y en la práctica estarán sujetas a todo tipo de procesos políticos que no tienen por qué generar consensos. Reconocer esta diversidad y pluralidad puede llevarnos a un enfoque pluralista. Este puede también tomar la forma del relativismo. La filosofía del siglo XXI y la teoría social, especialmente la llamada posmodernidad, han suscitado mucho escepticismo por lo que respecta a las respuestas que se presentan a sí mismas como universales, habiendo surgido en un contexto geográfico, histórico y cultural particular (por ejemplo, «Occidente») y en relación con intereses particulares y relaciones de poder. También se ha cuestionado si las políticas deberían buscar el consenso (véase, por ejemplo, Mouffe 2013); ¿el consenso es siempre deseable, o podría también tener algunos beneficios la lucha agonística sobre el futuro de la IA? Además, también está el problema relativo al *poder:* reflexionar sobre la ética en el mundo real significa reflexionar no solo sobre *qué* se necesita hacer en relación con la IA, sino también *quién* debería hacerlo y quién decidirá en la práctica sobre su futuro y, en consecuencia, sobre el futuro de nuestra sociedad. Considérense de nuevo los problemas del totalitarismo y del poder de las grandes empresas. Si rechazamos el totalitarismo y la plutocracia, ¿qué significa la toma democrática de decisiones con respecto a la IA? ¿Qué tipo de conocimiento sobre la IA necesitan políticos y ciudadanos? Si la IA y sus problemas potenciales apenas se entienden, nos enfrentamos al peligro de la tecnocracia, e incluso al riesgo de que no se desarrolle política alguna con respecto a la IA.

Sin embargo, como se muestra en el capítulo siguiente, al menos uno de los procesos políticos relevantes que han surgido recientemente parece querer llegar a tiempo. También es proactivo, busca el consenso, presenta un sorprendente grado de convergencia, proclama sin avergonzarse una universalidad, está basado en conocimiento especializado e incluso atiende a los ideales de la democracia, sirviendo tanto al bien como al interés común e

involucrando a las partes interesadas: hablo del diseño de políticas para la IA.

Capítulo 10

Políticas de actuación: propuestas

Qué se necesita hacer y otras preguntas que tienen que responder los responsables de diseñar las políticas de actuación

Dados los problemas éticos que suscita la IA, queda claro que algo debe hacerse. Es por ello que la mayoría de las iniciativas relacionadas con las políticas de actuación a este respecto incluyen una ética de la IA. Actualmente, dichas iniciativas son numerosas, lo cual debe aplaudirse. Sin embargo, no está claro qué es lo que debería hacerse ni cuáles son las líneas de actuación más apropiadas. Por ejemplo, no está especialmente claro qué debemos hacer para lidiar con la transparencia o el sesgo, dados la naturaleza de las tecnologías, el sesgo existente en la sociedad y las diferentes perspectivas que pueden adoptarse en torno a la justicia y la equidad. También hay muchas medidas posibles: una política puede implicar la regulación mediante leyes y directrices, como las regulaciones legales, pero también hay otras estrategias que pueden, o no, conectarse con las regulaciones legales, como medidas tecnológicas, códigos éticos y medidas educativas. Dentro de la regulación tampoco hay solo leyes, sino también estándares tales como las normas ISO. Además, hay otros tipos de cuestiones que necesitan responderse proponiendo políticas de actuación: no solo *qué* debe hacerse, sino también *por qué*, *cuándo, cuánto*, *por parte de quién,* y cuáles son la *naturaleza, extensión* y *urgencia del problema*.

Es importante justificar las medidas propuestas. Por ejemplo, una propuesta puede basarse en principios de derechos humanos para reducir el sesgo en la toma de decisiones algorítmica. Segundo, en respuesta al desarrollo de la tecnología, la política de actuación llega a menudo demasiado tarde, cuando la tecnología ya está integrada en la sociedad. En vez de lo anterior,

se podría intentar diseñar las políticas de actuación antes de que la tecnología esté completamente desarrollada y en uso. En el caso de la IA esto es aún posible en cierto grado, a pesar de que buena parte de la IA ya esté funcionando fuera. La dimensión temporal también es relevante por lo que respecta al alcance de la política que se propone: ¿está concebida para los próximos cinco o diez años, o como un marco a largo plazo? Aquí tenemos que elegir. Por ejemplo, se podrían desechar las predicciones a largo plazo y centrarse en el futuro cercano, como se hace en muchas propuestas, o se podría ofrecer una visión del futuro de la humanidad. Tercero, no todo el mundo está de acuerdo en que para resolver los problemas se necesiten muchas medidas nuevas. Algunas personas y organizaciones han argumentado que la legislación actual es suficiente para lidiar con la IA. Si este es el caso, entonces parece que los legisladores no tienen mucho trabajo por delante, pero sí aquellos que interpretan la ley y quienes desarrollan la IA. Otros piensan que fundamentalmente necesitamos repensar la sociedad y sus instituciones, incluyendo nuestros sistemas legales, para poder lidiar con los problemas subyacentes y preparar a las generaciones futuras. Cuarto, una propuesta de política de actuación debería dejar claro a quién le corresponde llevarla a cabo. Puede que no sea un solo actor, sino más de uno. Como hemos visto, hay muchas manos involucradas en cualquier acción tecnológica. Esto plantea la cuestión de cómo distribuir la responsabilidad respecto de las políticas tecnológicas y el cambio: ¿corre principalmente de parte de los gobiernos llevar a cabo la acción, o deberían, por ejemplo, las empresas y la industria desarrollar sus propias líneas de acción para asegurar una IA ética? Por lo que atañe a las empresas, ¿deberían estar dirigidas solo a las grandes corporaciones o también a los negocios pequeños y medianos? ¿Y cuál es el papel de los científicos, informáticos e ingenieros? ¿Cuál el de los ciudadanos?

Quinto, responder a lo que debe hacerse, cuánto debe hacerse y otras cuestiones por el estilo, depende principalmente de cómo se definan la naturaleza, la extensión y la urgencia del problema mismo. Por ejemplo, se da una tendencia en las políticas relacionadas con la tecnología (y, ciertamente, en la ética de la IA) que consiste en ver problemas nuevos en todas partes. Sin embargo, en el capítulo anterior hemos visto que muchos problemas

pueden no ser exclusivos de las nuevas tecnologías, sino que quizás hayan existido durante mucho tiempo. Además, como ya quedó claro cuando discutíamos el sesgo, lo que propongamos hacer dependerá de cómo definamos el problema: ¿es un problema de justicia?, y si es así, ¿qué tipo de justicia es la que se ve amenazada? La definición determinará las medidas propuestas. Por ejemplo, si se proponen medidas de acción afirmativa, entonces estarán basadas en una definición del problema particular. Finalmente, también desempeña un papel la definición misma de la IA, que siempre es discutible y que tiene importancia en lo que se refiere al alcance de la política. Por ejemplo, ¿es posible y deseable distinguir claramente entre IA y algoritmos autónomos inteligentes, o entre IA y tecnologías de la automatización? Todas estas cuestiones convierten el diseño de políticas en una empresa controvertida. Y, ciertamente, puede que encontremos muchos desacuerdos y tensiones, por ejemplo, sobre cuánta legislación es necesaria, cuáles son los principios en los que esta debe basarse exactamente, el problema de si la ética debería ponderarse con otros asuntos (por ejemplo, competitividad de las empresas y la economía). Sin embargo, si consideramos los documentos en que se plasman las políticas de actuación reales, también encontraremos un notable grado de convergencia.

Principios éticos y justificaciones

La intuición ampliamente compartida de que es urgente e importante lidiar con los desafíos éticos y sociales planteados por la IA ha generado una avalancha de iniciativas y documentos relativos a políticas de actuación que no solo identifican algunos problemas éticos en la IA, sino que también buscan facilitar una orientación normativa para dichas políticas. Han sido muchos los actores que han propuesto políticas de actuación respecto de la IA desde una perspectiva ética, incluyendo gobiernos y cuerpos gubernamentales tales como comités éticos nacionales, compañías tecnológicas como Google, ingenieros y sus organizaciones profesionales como IEEE (N. del T.: Institute of Electrical and Electronics Engineers, en castellano, Instituto de ingenieros eléctricos y electrónicos), organizaciones interguber-

namentales como la UE, actores no gubernamentales sin ánimo de lucro e investigadores.

La intuición ampliamente compartida de que es urgente e importante lidiar con los desafíos éticos y sociales planteados por la IA ha generado una avalancha de iniciativas y documentos relativos a políticas de actuación.

Si revisamos algunas iniciativas y propuestas recientes, observaremos que la mayoría de los documentos comienzan con la justificación de la política a seguir, disponiendo principios y haciendo recomendaciones con respecto a los problemas éticos identificados. Como veremos, estos *problemas y principios* son muy similares. Las iniciativas normalmente se basan en principios generales éticos y principios de códigos éticos profesionales. Revisaré ahora algunas propuestas.

La mayoría de propuestas rechazan el escenario de ciencia ficción en el que una máquina superinteligente asume el control. Por ejemplo, bajo la presidencia de Obama, el gobierno estadounidense publicó el informe «Preparación para el futuro de la inteligencia artificial», que afirma explícitamente que las preocupaciones a largo plazo sobre una IA general superinteligente «deberían tener poco impacto en la política actual» (Executive Office of the President 2016, 8). En su lugar, el informe discute los problemas actuales y del futuro cercano planteados por el aprendizaje automático, como el sesgo y el problema de que incluso los desarrolladores puedan no comprender su sistema lo suficientemente bien como para evitar ciertas consecuencias. El informe subraya que la IA es buena para la innovación y el crecimiento económico, y hace hincapié en la autorregulación, pero dice que el gobierno estadounidense puede monitorizar la seguridad y la legitimidad de las aplicaciones y, si fuera necesario, adaptar los marcos regulatorios.

Muchos países de Europa disponen ya de estrategias para la IA que incluyen un componente ético. La «IA explicable» es una meta compartida por muchos diseñadores de políticas de actuación. La Cámara de los Comunes del Reino Unido (2018) afirma que la transparencia y el derecho a la explicación son claves para la responsabilidad algorítmica, y que las industrias y los reguladores deben abordar el problema de los sesgos en procesos algorítmicos de toma de decisión. El selecto comité de la Cámara de los Lores británica sobre la IA también examina sus implicaciones éticas. En Francia, el informe Villani propone trabajar en dirección hacia una «IA significativa» que no refuerce los problemas de exclusión, incremente la desi-

gualdad o conduzca a una sociedad en la que estemos gobernados por algoritmos de caja negra: la IA debería ser explicable y respetuosa con el medio ambiente (Villani 2018). Austria ha conformado recientemente un consejo asesor nacional dedicado a la robótica y a la IA que hace recomendaciones políticas basadas en los derechos humanos, la justicia y la equidad, la inclusividad y la solidaridad, la democracia y la participación, la no discriminación, la responsabilidad y otros valores similares. Su documentación técnica también recomienda una IA explicable y dice explícitamente que la responsabilidad sigue residiendo en los humanos: las IAs no pueden ser moralmente responsables (ACRAI 2018). Las organizaciones y los congresos internacionales también están muy activos. Por ejemplo, la Conferencia internacional de comisarios de protección de datos y privacidad ha publicado una declaración sobre la ética y la protección de datos en IA que incluye principios de justicia, responsabilidad, transparencia e inteligibilidad, diseño responsable y privacidad en virtud del diseño (un concepto que exige tener en cuenta la privacidad a lo largo del proceso completo de ingeniería), empoderamiento de individuos y reducción y mitigación de los sesgos o la discriminación (ICDPPC 2018).

Algunos diseñadores de políticas de actuación en IA enmarcan su objetivo refiriéndose a la «IA fiable». La Comisión Europea, por ejemplo, sin duda uno de los principales actores globales en el área del diseño de políticas para la IA, hace mucho hincapié en el término. En abril de 2018, conformó un Grupo de Expertos de Alto Nivel en inteligencia artificial para crear un nuevo conjunto de directrices sobre esta; en diciembre de 2018 el grupo publicó un borrador con directrices éticas que reclaman un enfoque antropocéntrico para la IA y el desarrollo de una IA fiable, que respete los derechos fundamentales y principios éticos. Los derechos mencionados son la dignidad humana, la libertad del individuo, el respeto por la democracia, la justicia y el estado de derecho, así como por los derechos de los ciudadanos. Los principios éticos son la beneficencia (hacer el bien) y no causar daño, la autonomía (preservar la capacidad de actuación humana), la justicia (ser justo) y la explicabilidad (operar de forma transparente). Estos principios nos son familiares en bioética, pero el documento añade la explicabilidad e incluye interpretaciones que subrayan los problemas éticos específi-

cos planteados por la IA. Por ejemplo, el principio de no causar daño es interpretado como el requisito de que los algoritmos de la IA eviten siempre la discriminación, la manipulación y la clasificación negativa, y protejan a los grupos vulnerables como niños e inmigrantes. El principio de justicia es interpretado como la inclusión de la exigencia de que los desarrolladores e implementadores de la IA se aseguren de que los individuos y los grupos minoritarios se mantengan libres de sesgo. El principio de explicabilidad se presenta como la exigencia de que los sistemas sean auditables, así como «comprensibles e inteligibles para los seres humanos a distintos niveles de comprensión y experiencia» (European Commission AI HLEG 2018, 10). La versión final, publicada en abril de 2019, especifica que la explicabilidad no tiene únicamente que ver con explicar el proceso técnico, sino también las decisiones humanas relacionadas (European Commission AI HLEG 2019, 18).

Anteriormente, otro organismo asesor de la UE, el grupo europeo para la ética en la ciencia y las nuevas tecnologías (EGE), publicó una declaración sobre IA, robótica y sistemas autónomos, proponiendo los principios de dignidad humana, autonomía, responsabilidad, justicia, equidad, solidaridad, democracia, estado de derecho y responsabilidad, seguridad, protección de datos y privacidad, y sostenibilidad. El principio de la dignidad humana implica que a la gente se la tiene que hacer consciente de si está interactuando con una máquina o con otro ser humano (EGE 2018). Nótese también que la UE ya tiene regulaciones preparadas que resultan relevantes para el desarrollo y uso de la IA. El Reglamento General de Protección de Datos (GDPR por sus siglas en inglés), que fue promulgado en mayo de 2018, pretende proteger y empoderar a los ciudadanos europeos en lo que respecta a la privacidad de datos. Incluye principios como el derecho al olvido (la persona a la que conciernen los datos puede pedir borrar sus datos personales y detener su procesamiento posterior) y la privacidad en virtud del diseño. También otorga a los sujetos a los que conciernen los datos el derecho a acceder a «información significativa sobre la lógica involucrada» en la toma de decisiones automatizada e información sobre las «consecuencias previstas» de tal procesamiento (Parlamento Europeo y Consejo Europeo 2016). La diferencia con los documentos normativos es que, aquí, estos

principios son requerimientos legales. Es legislación que está en vigor: las organizaciones que infrinjan el GDPR pueden ser sancionadas. Sin embargo, se ha cuestionado si las disposiciones del GDPR equivalen a un derecho completo a la explicación de la decisión (Digital Europe 2018) y, de forma más general, si ofrecen protección suficiente contra los riesgos de la toma de decisiones automatizada (Wachter, Mittelstadt y Floridi 2017). El GDPR proporciona un derecho a ser informado sobre las decisiones tomadas de forma automatizada pero no parece exigir una explicación sobre las motivaciones de ninguna decisión en particular. Esta es también una preocupación por lo que se refiere a la toma de decisiones en la esfera legal. Un estudio del Consejo Europeo, basado en el trabajo de un comité de expertos en derechos humanos, exigió que los individuos tuvieran derecho a un juicio justo en términos comprensibles para ellos (Yeung 2018).

Los debates legales son, por supuesto, muy relevantes para las discusiones sobre la ética de la IA y las políticas de actuación respecto de la IA. Turner (2019) ha examinado ciertas comparaciones con animales (cómo son y cómo son tratados por la ley y si tienen derechos) y revisado una serie de instrumentos legales relacionados con lo que una equiparación en este sentido podría significar para la IA. Por ejemplo, cuando se ha infligido un daño, la negligencia se produce si una persona tenía el deber de proteger a la parte dañada o prevenir dicho daño, incluso si este no ha sido intencionado. Esto se podría aplicar al diseñador o entrenador de la IA. Pero ¿cómo de fácil es prever las consecuencias de la IA? En derecho penal, por contra, se necesita la intención de hacer daño. Pero este no es normalmente el caso de la IA. Por otro lado, la responsabilidad respecto de los productos no es algo que concierna a un fallo de los individuos, sino que es la empresa que produjo la tecnología la que paga por los daños, independientemente de a quién corresponda exactamente la responsabilidad. Esta podría ser una posible solución para la responsabilidad legal de la IA. Las leyes de propiedad intelectual, tales como el copyright y las patentes, son también relevantes, y han hecho surgir discusiones sobre la «personalidad legal» de las IAs, una ficción legal que, sin embargo, ya se está aplicando como instrumento en empresas y distintas organizaciones. ¿Debería también aplicarse a la IA? En una controvertida resolución de 2017, el Parlamento Europeo sugirió que otorgar a

los robots con la más sofisticada autonomía el estatus de personas electrónicas era una posible solución para la cuestión de la responsabilidad legal: una idea que no asumió la Comisión Europea en su estrategia de 2018 para la IA . Otros se han opuesto vehementemente a la idea misma de otorgar derechos y personalidad a las máquinas, argumentando, por ejemplo, que entonces sería difícil, si no imposible, exigir responsabilidades a nadie, puesto que habría quien explotaría el concepto con fines egoístas (Bryson, Diamantis y Grant 2017). También está el famoso caso de Sophia, un robot al que Arabia Saudí ortorgó la «ciudadanía» en 2017. Un caso así hace surgir de nuevo la cuestión del estatus moral de los robots y las IAs (véase el capítulo 4).

También se han propuesto políticas de actuación con respecto a la IA en regiones distintas de América del Norte y Europa. China, por ejemplo, tiene una estrategia nacional para la IA. Su plan de desarrollo reconoce que la IA es una tecnología disruptiva que puede afectar a la estabilidad social, repercutir en la ley y en la ética social, violar la privacidad personal y dar lugar a riesgos en seguridad; el plan, por lo tanto, recomienda fortalecer la prevención avanzada y minimizar dicho riesgo (Consejo de Estado de China 2017). Algunos actores en Occidente hablan en términos de narrativa competitiva: temen que China les supere, o incluso que nos estemos aproximando a una nueva guerra mundial. Otros intentan *aprender* de la estrategia de China. Los investigadores también se preguntan cómo y de qué manera diferentes culturas abordan este problema desde sus respectivos puntos de vista. La propia investigación en IA puede llevarnos a adoptar una perspectiva más intercultural o proclive a la comparación sobre la ética en la IA, por ejemplo, si tenemos en cuenta las diferencias entre culturas individualistas y colectivistas en lo que se refiere a los dilemas morales (Awad *et al.* 2018). Esto podría hacer surgir problemas para la ética en la IA si pretende ser universal. También se puede explorar cómo las narrativas sobre la IA en China o en Japón, por ejemplo, difieren de las occidentales. Con todo, a pesar de las diferencias culturales, lo cierto es que las políticas éticas con respcto a la IA resultan tremendamente similares. Si bien el plan de China pone más énfasis en la estabilidad social y en el bien colectivo, los riesgos

éticos identificados y los principios mencionados no son tan diferentes de aquellos propuestos por los países occidentales.

Pero, como mencioné anteriormente, la política de actuación sobre la ética de la IA no está totalmente limitada a los gobiernos y a sus comités y corporaciones. Los académicos también han emprendido iniciativas. Por ejemplo, la Declaración de IA responsable de Montreal fue propuesta por la Universidad de Montreal e incluyó la consulta a ciudadanos, expertos y otras partes interesadas. Afirma que el desarrollo de la IA debería promover el bienestar de todas las criaturas sintientes y la autonomía de los seres humanos, eliminar todo tipo de discriminación, respetar la privacidad personal, protegernos de la propaganda y la manipulación, promover el debate democrático y hacer responsables a distintos actores a la hora de trabajar contra los riesgos de la IA (Universidad de Montreal 2017). Otros investigadores han propuesto los principios de beneficencia, no maleficencia, autonomía, justicia y explicabilidad (Floridi *et al.* 2018). Universidades como Cambridge y Stanford trabajan en la ética de la IA, a menudo desde la perspectiva de la ética aplicada. También las personas que trabajan en el campo de la ética profesional están haciendo un trabajo provechoso. Por ejemplo, el Centro Markkula para la Ética Aplicada de la Universidad de Santa Clara ha ofrecido una serie de teorías éticas como caja de herramientas para la práctica tecnológica y de ingeniería, que puede también ser útil a la ética de la IA . Además, los filósofos de la tecnología han mostrado recientemente un notable interés.

También encontramos iniciativas relacionadas con la ética de la IA en el mundo corporativo. Por ejemplo, el Partnership on AI (Asociación sobre la IA) incluye empresas tales como DeepMind, IBM, Intel, Amazon, Apple, Sony y Facebook . Muchas empresas reconocen la necesidad de una IA ética. Por ejemplo, Google ha hecho pública una serie de principios éticos para la IA: ofrecer beneficios sociales, evitar crear o reforzar sesgos injustos, reforzar la seguridad, promover la excelencia científica y limitar aplicaciones potencialmente dañinas o abusivas como armas o tecnologías que violen los principios de la ley internacional y los derechos humanos . Microsoft habla de la «IA para el Bien» y propone los principios de justicia, fiabilidad e inocuidad, privacidad y seguridad, inclusividad, transparencia y

responsabilidad . Accenture ha propuesto principios universales para la ética relativa a los datos, incluyendo el respeto a las personas que se encuentran tras ellos, la privacidad, la inclusión y la transparencia . Y, a pesar de que en los documentos corporativos el énfasis suele ponerse en la autorregulación, algunas empresas reconocen la necesidad de regulación externa. Por ejemplo, el CEO de Apple Tim Cook ha manifestado que la regulación tecnológica es inevitable para asegurar la privacidad, porque el libre mercado no está funcionando . Sin embargo, existe el debate sobre si se requiere una nueva regulación. Algunos apoyan el camino de la regulación, incluyendo nuevas leyes. California ya ha propuesto un proyecto de ley que exige que se conozca la identidad real de los *bots:* es ilegal utilizar un *bot* si se hace de una forma que engañe a la otra persona sobre su identidad artificial . Otros toman una posición más conservadora. Digital Europe (2018), que representa a la industria digital de Europa, ha argumentado que el marco legal actual está preparado para abordar problemas relacionados con la IA, incluyendo el sesgo y la discriminación, pero que para construir confianza, transparencia, explicabilidad e interpretabilidad es importante lo siguiente: la gente y las empresas deben entender cuándo y cómo se usan los algoritmos para tomar decisiones, y necesitamos proporcionar información significativa y facilitar la interpretación de las decisiones basadas en algoritmos.

A pesar de las diferencias culturales, las políticas éticas con respecto a la IA resultan tremendamente parecidas en distintos países.

Ciertos actores sin ánimo de lucro también desempeñan un papel relevante. Por ejemplo, el movimiento internacional llamado Campaign to Stop Killer Robots (Campaña contra los robots asesinos) ha llamado la atención acerca de numerosos problemas éticos en relación con las aplicaciones militares de la IA . Desde el lado transhumanista, se han planteado los principios Asilomar de IA, acordados por participantes académicos y del sector industrial en una conferencia convocada por el Future of Life Institute (Instituto para el futuro de la vida) (Max Tegmark y otros). El objetivo global es mantener la IA como algo beneficioso y respetar principios y valores éticos como la seguridad, la transparencia, la responsabilidad, la orientación hacia los valores, la privacidad y el control humano . También hay organizaciones profesionales que trabajan en políticas de actuación respecto de la IA. El Instituto de Ingenieros Eléctricos y Electrónicos (IEEE), que afirma ser la organización técnica más grande del mundo, ha presentado una iniciativa global sobre la ética de sistemas autónomos e inteligentes. Después de haber tenido lugar debates entre expertos, la iniciativa ha producido un documento con un planteamiento a favor del «diseño éticamente alineado», proponiendo que el diseño, desarrollo e implementación de estas tecnologías esté siempre guiado por los principios generales de los derechos humanos, el bienestar, la responsabilidad, la transparencia y la conciencia del uso indebido. Implementar la ética en estándares técnicos globales puede ser una forma efectiva de contribuir al desarrollo de una IA ética.

Las soluciones tecnológicas y la cuestión de los métodos y de la operacionalización

La iniciativa global del IEEE muestra que, en términos de medidas, algunas propuestas se centran en soluciones tecnológicas. Por ejemplo, como mencioné en el capítulo anterior, ciertos investigadores han reclamado abrir la caja negra, es decir, una inteligencia artificial explicable. Hay buenas razones para defender esto: explicar los motivos que están detrás de la decisión que uno toma no es solo éticamente necesario, sino también un aspecto

importante de la inteligencia humana (Samek, Wiegand y Müller 2017). La idea de una IA explicable o transparente pasa, pues, por que las acciones y decisiones tomadas por las IAs sean fácilmente comprensibles. Como hemos visto, esta idea es difícil de implementar en el caso del aprendizaje automático que se vale de redes neuronales (Goebel *et al.* 2018). Pero las políticas de actuación pueden, por supuesto, apoyar la investigación en esta dirección.

En general, integrar a la ética en el diseño de nuevas tecnologías es una idea excelente. Ideas como la del ética en el diseño o la del diseño sensible al valor, que tienen su propia historia , pueden ayudarnos a crear una IA que conduzca a una mayor responsabilidad y transparencia. Por ejemplo, la ética incorporada en el diseño puede incluir el requisito de que la trazabilidad esté asegurada en todas las fases (Dignum *et al.* 2018), contribuyendo, así, a la rendición de cuentas por parte de la IA. La idea de la trazabilidad se puede tomar literalmente, en tanto que sería posible grabar los datos sobre el comportamiento del sistema. Winfield y Jirotka (2017) han pedido implementar una «caja negra ética» en robots y sistemas autónomos, que grabe lo que el robot hace (datos provenientes de los sensores y del estado «interno» del sistema), como sucede con las cajas negras instaladas en los aviones. Esta idea también podría aplicarse a IAs autónomas: cuando algo salga mal, estos datos podrían ayudarnos a explicar qué es exactamente lo que ha fallado. Ciertamente, y como bien han observado los investigadores, podemos aprender mucho de la industria aeronáutica, que está altamente regulada y tiene procedimientos rigurosos de certificación de la seguridad y procedimientos visibles de investigación de accidentes. ¿Podría instalarse en la IA una infraestructura de regulación y seguridad similar? Por hacer la comparación con otro sector del transporte, la industria automovilística también ha propuesto certificaciones o algún tipo de «carnet de conducir» para vehículos autónomos con IA . Algunos investigadores van más allá y tratan de crear máquinas morales que permitan una «ética de las máquinas» en el sentido de que las máquinas mismas sean capaces de tomar sus decisiones. Otros argumentan que esa es una idea peligrosa y que la decisión debe reservarse a los humanos, que es imposible crear agentes completamente éticos, que no hay necesidad de que las máquinas sean agentes éticos comple-

tos y que es suficiente con que sean seguras y respetuosas con la ley (Yampolskiy 2013); o que podría haber formas de «moralidad funcional» (Wallach y Allen 2009) que, sin representar una moralidad completa, permitan conseguir máquinas relativamente morales. Esta discusión, que apunta de nuevo a la cuestión del estatus moral, es relevante, por ejemplo, en el caso de los coches autónomos: ¿hasta qué punto es necesario, posible y deseable incorporar la ética a estos coches?, ¿qué tipo de ética debería ser esta y cómo deberíamos implementarla técnicamente?

Los encargados de diseñar políticas de actuación tienden a respaldar muchas de estas guías en la investigación e innovación en IA, como la IA explicable y, más generalmente, la ética integrada en el diseño. Por ejemplo, junto a métodos no técnicos, como la regulación, la estandarización, la educación, el diálogo entre las partes interesadas y los equipos de diseño inclusivo, el informe del High-Level Expert Group (Grupo de Expertos de Alto Nivel) menciona una cantidad de *métodos* técnicos que incluyen a la ética y al estado de derecho en el diseño, arquitecturas para una IA fiable, pruebas y validaciones, trazabilidad y auditabilidad, y explicación. Por ejemplo, la ética incorporada en el diseño puede incluir la privacidad. El informe también menciona algunas formas en las que se puede *operacionalizar* la IA fiable, como la trazabilidad, en tanto que permite contribuir a la transparencia: en el caso de IA basada en reglas debería dejarse claro cómo se ha construido el modelo, y en el caso de la IA basada en aprendizaje, el método de entrenamiento del algoritmo, incluyendo cómo se recogieron y seleccionaron los datos. Esto aseguraría que el sistema de IA es auditable, especialmente en situaciones críticas (Comisión Europea de la IA HLEG 2019).

Ideas como la de la ética en el diseño o la del diseño sensible al valor pueden ayudarnos a crear una IA que conduzca a una mayor, responsabilidad y transparencia.

La cuestión de los métodos y la operacionalización es crucial: una cosa es hacer una lista con una cierta cantidad de principios éticos y otra bien distinta averiguar cómo implementarlos en la práctica. Incluso conceptos tales como la privacidad en virtud del diseño, que se supone que están más cerca del proceso de desarrollo y de la ingeniería, son formulados normalmente de manera abstracta y general: *no queda claro qué deberíamos hacer exactamente.* Esto nos conduce al siguiente capítulo, en el que nos aventuraremos brevemente en el debate en torno a ciertos desafíos a los que se enfrentan las políticas de actuación respecto de la ética de la IA.

Capítulo 11

Desafíos a los que se enfrentan los encargados del desarrollo de políticas de actuación

Ética proactiva: innovación responsable y valores integrados en el diseño

Quizás no sea de extrañar, pero la política de actuación respecto de la ética de la IA se enfrenta a numerosos desafíos. Hemos visto que algunas propuestas normativas se apoyan en una visión de la ética de la IA que es *proactiva,* es decir, que plantean que es necesario tener en cuenta la ética desde el principio en el desarrollo tecnológico. La idea es evitar problemas éticos y sociales generados por la IA que puedan ser difíciles de resolver una vez esta haya aparecido. Esta cuestión está en sintonía con ideas como la innovación responsable, la integración de valores en el diseño y otras propuestas similares de los últimos años. Deja a un lado el problema de tener que lidiar con los efectos negativos de la tecnología que ya se está usando ampliamente en favor de asumir responsabilidades por las tecnologías que se están desarrollando hoy en dia.

Sin embargo, no es fácil prever las consecuencias que podrían tener las nuevas tecnologías en fase de diseño. Una forma de mitigar este problema es construir escenarios hipotéticos en torno a futuros conflictos de índole ética. Existen distintos métodos para practicar la investigación y la innovación éticamente (Reijers *et al.* 2018), uno de los cuales pasa, además de por estudiar y valorar el impacto de las narrativas actuales de la IA (Royal Society 2018), por crear narrativas nuevas, más específicas, en torno a sus aplicaciones concretas.

PRAGMATISMO Y *BOTTOM-UP:* ¿CÓMO LLEVARLOS A LA PRÁCTICA?

La innovación responsable no consiste solamente en integrar la ética en el diseño, sino que también requiere tener en consideración las opiniones e intereses de las distintas partes. La gestión inclusiva conlleva una implicación mayor de estas últimas, más debate público y una intervención social temprana en la investigación y la innovación (Von Schomberg 2011). Esto puede implicar, por ejemplo, organizar grupos de discusión y emplear distintas técnicas para ver lo que la gente piensa de la tecnología.

Este enfoque de la innovación responsable estructurado de forma ascendente *(bottom-up,* en inglés) está de alguna forma en conflicto con el enfoque que se da a la ética aplicada en la mayoría de propuestas normativas, que, más bien, suelen estar estructuradas de forma descendente *(top-down)* y son tremendamente abstractas. Hay que tener en cuenta, en primer lugar, que las políticas suelen ser diseñadas por especialistas, sin que otras partes interesadas aporten gran cosa; en segundo lugar, incluso si dichas políticas apoyan principios como la ética en virtud del diseño, tienden a seguir siendo demasiado vagas sobre qué significa aplicar estos principios en la práctica. Para hacer que las políticas que atañen a la IA funcionen, sigue siendo necesario tender un puente entre, por un lado, estos principios abstractos y de alto nivel ético y legal y, por otro, las prácticas de desarrollo y uso de la tecnología en contextos concretos, las tecnologías en sí y las voces de aquellos que son parte de estas prácticas y trabajan en estos contextos. ¿Se puede y se debe hacer más en las primeras fases del diseño de las políticas? En cualquier caso es seguro que hace falta trabajar más, tanto en el «cómo» como en el «qué», así como en los métodos, procedimientos e instituciones que resultan necesarios para llevar a la práctica los esfuerzos dedicados a la ética de la IA. Debemos prestar más atención al *proceso.*

Con respecto a la pregunta del «quién», necesitamos más espacio para procedimientos ascendentes que coexistan con los descendentes, es decir, escuchar más a los investigadores y profesionales que trabajan con IA en la práctica y, ciertamente, a las personas potencialmente perjudicadas por ella. Si apoyamos el ideal de la democracia y si este concepto implica la inclusividad y la participación en la toma de decisiones sobre el futuro de nuestras

sociedades, entonces escuchar la voz de todas las partes interesadas no es algo opcional, sino necesario ética y políticamente. Aunque hay encargados de diseñar políticas de actuación que se relacionan de algún modo con dichas partes (por ejemplo, la Comisión Europea tiene su AI Alliance [Alianza de la IA]) , sigue sin estar claro si sus esfuerzos tienen un impacto verdadero para los desarrolladores, los usuarios finales de la tecnología y, sobre todo, aquellos que tengan que correr con los riesgos y vivir con sus consecuencias negativas. ¿Cómo de democráticas y participativas son en realidad la toma de decisiones y las políticas concernientes a la IA?

La innovación responsable no consiste solamente en integrar la ética en el diseño, sino que también requiere tener en consideración las opiniones e intereses de las distintas partes interesadas.

El ideal de la democracia también se encuentra en peligro por el hecho de que el poder está concentrado en manos de un relativamente pequeño número de grandes corporaciones. Paul Nemitz (2018) ha señalado que tal acumulación de poder digital en manos de unos pocos es problemática: si un puñado de empresas ejercen el poder no solo sobre los individuos (pues mediante la catalogación de nuestros perfiles centralizan el poder) sino también sobre las infraestructuras para la democracia, entonces, y aunque traten de contribuir a la IA ética con sus mejores intenciones, en realidad le están poniendo barreras. Es, por tanto, necesario regular y establecer límites, y asegurarnos de que estas empresas no determinan por sí mismas las normas. Murrah Shanahan también ha señalado la «tendencia del poder, la riqueza y los recursos de auto-perpetuarse, de concentrarse en manos de unos pocos» (2015, 166), la cual hace difícil lograr una sociedad más equitativa. También hace a las personas más vulnerables a todo tipo de riesgos, incluyendo la explotación y violación de la privacidad, por ejemplo, a la que un estudio del Consejo de Europa se refiere como «el aterrador efecto de la reutilización de datos para propósitos distintos de los originales» (Yeung 2018, 33).

Si comparamos la situación con la de la política medioambiental, cabe ser pesimistas sobre la posibilidad de que los países tomen medidas efectivas y emprendan acciones colaborativas con respecto a la ética de la IA. Considérense, por ejemplo, los procesos políticos que rodean a la cuestión del cambio climático en Estados Unidos, donde, a veces, incluso se niegan el *problema* mismo del calentamiento global y el cambio climático y donde poderosas fuerzas políticas trabajan en contra de cualquier acción; o el más bien limitado éxito que han tenido las cumbres internacionales sobre el clima para alcanzar acuerdos en torno a una política climática común y efectiva. Puede que aquellos que luchan por emprender acciones globales ante los problemas éticos y sociales planteados por la IA se enfrenten a dificultades similares. A menudo prevalecen intereses distintos del bien público, y la acción política intergubernamental en torno a las nuevas tecnologías digitales, incluyendo la IA, es demasiado escasa. Una excepción está en el interés global en prohibir armas letales autónomas, que tienen componentes de IA,

pero no deja de ser una rareza, y tampoco es algo que compartan todos los países (sigue siendo controvertido en los EEUU, por ejemplo).

Además, y aunque sean bien intencionadas, la ética en virtud del diseño y la innovación responsable tienen sus propias limitaciones. En primer lugar, los métodos como el diseño sensible a los valores presuponen que nosotros mismos somos capaces de determinar con precisión nuestros valores, y los esfuerzos por construir máquinas morales, que podemos acotar completamente nuestra ética. Pero este no es siempre el caso: nuestra ética del día a día en absoluto tiene por qué responder a un sistema perfectamente articulado. A veces reaccionamos a problemas éticos sin ser capaces de justificar nuestra respuesta (Boddington 2017). Tomando prestado un término de Wittgenstein: nuestra ética no está solo corporeizada sino también integrada en una *forma de vida.* Está profundamente vinculada a la manera en que hacemos las cosas como seres encarnados y sociales, y también como sociedades y culturas. Esto impone límites a la hora de determinar la ética y el razonamiento moral. Supone un problema para el proyecto de desarrollar máquinas morales y desafía la hipótesis de que la ética y la democracia pueden llegar a ser *completamente* deliberativas. También es un problema para los diseñadores de políticas relacionadas con la IA que piensan que la ética de esta puede abordarse en su totalidad mediante una lista de principios o mediante métodos legales y técnicos específicos. No cabe duda de que necesitamos métodos, procedimientos y operaciones, pero no son suficientes por sí solos: la ética no funciona como una máquina, ni tampoco los diseños, ni la innovación responsable.

En segundo lugar, estos enfoques también pueden poner barreras a la ética cuando sea necesario frenar el desarrollo de una tecnología. En la práctica suelen funcionar como una suerte de aceite que ayuda a lubricar la maquinaria de la innovación para incrementar los beneficios y asegurar la aceptabilidad de la tecnología, lo cual no tiene por qué ser malo. ¿Pero qué pasa si los principios éticos implican que la tecnología, o una aplicación concreta de la misma, deben cesarse o congelarse? Crawford y Calo (2016) han defendido que las herramientas del diseño sensible a los valores y de la innovación responsable funcionan bajo el supuesto de que la tecnología va a desarrollarse; resultan menos útiles si lo que debe decidirse es si se desarro-

lla o no. Por ejemplo, en el caso de IAs avanzadas, como las nuevas aplicaciones de aprendizaje automático, puede ser que la tecnología no sea aún fiable, o tenga serios inconvenientes éticos que impliquen que al menos algunas de sus aplicaciones no se deberían llevar a la práctica (por el momento). Independientemente de que optemos por detener o no este tipo de desarrollos, lo fundamental es que dispongamos, al menos, de margen para plantearnos la cuestión y decidir. Si dicho margen no existe, la innovación responsable no es más que una tapadera para seguir trabajando como hasta ahora.

Hacia una ética positiva

Dicho esto, la ética de la IA, en general, no implica únicamente prohibiciones (Boddington 2017). Otra barrera que impide la ética en la práctica reside en que muchos actores del campo de la IA, como las empresas y los investigadores técnicos, piensan aún en ella como una limitación, como algo negativo. La idea no es del todo errónea: a menudo la ética tiene que restringir, limitar, tiene que decir que algo es inaceptable. Si nos tomamos la ética de la IA seriamente e implementamos sus recomendaciones, puede que tengamos que enfrentarnos a contrapartidas, sobre todo a corto plazo. La ética puede tener un coste de dinero, tiempo y energía. Sin embargo, al reducir los riesgos, la ética y la innovación responsable apoyan el desarrollo sostenible a largo plazo de las empresas y de la sociedad. Sigue siendo un reto convencer a todos los actores en el campo de la IA, incluyendo a los que desarrollan políticas de actuación, de que este es ciertamente el caso. Obsérvese que tampoco las políticas y la regulación tienen solo que ver con prohibir cosas o ponerlas más difíciles: también pueden servir de apoyo, ofreciendo incentivos, por ejemplo.

Adicionalmente, además de una ética negativa que imponga límites, también necesitamos hacer explícita y elaborar una ética *positiva* con el objetivo de desarrollar visiones buenas de la vida y de la sociedad. Aunque algunos principios éticos propuestos anteriormente ya insinúan una visión de este tipo, sigue siendo un desafío redirigir la discusión en esta dirección. Como argumenté anteriormente, los problemas éticos relativos a la IA no

afectan solo a la tecnología, sino también a las vidas humanas y su prosperidad, el futuro de la sociedad, y quizás también a los seres no humanos, el medio ambiente y el del planeta (véase el siguiente capítulo). Los debates sobre la ética en la IA y las políticas relativas a ella nos conducen, de nuevo, a las grandes preguntas que tenemos que hacernos: como individuos, como sociedades y, quizás, como humanidad. Los filósofos pueden ayudar a reflexionar sobre estas cuestiones. Para los encargados de diseñar políticas de actuación, el reto consiste en poner en juego una visión amplia del futuro tecnológico que incluya ideas sobre lo que es importante, significativo y valioso. Aunque, en general, las democracias liberales están creadas para dejar estas cuestiones en manos de los individuos y se supone que son poco intervencionistas en torno a la determinación de temas fundamentales, como el de que en qué consiste una buena vida (una innovación política que ha evitado, al menos, ciertos tipos de guerra y ha contribuido a la estabilidad y a la prosperidad), a la vista de los desafíos éticos y políticos a los que nos enfrentamos sería irresponsable rechazar intervenir en los problemas éticos de mayor calado. Las políticas de actuación, incluyendo las que conciernen a la IA, deberían ocuparse también de la ética positiva.

La ética de IA no implica únicamente prohibiciones; también necesitamos una ética *positiva* con el objetivo de desarrollar visiones buenas de la vida y de la sociedad.

Sin embargo, los diseñadores de políticas de actuación para la IA no deben actuar por cuenta propia y adoptando el rol del filósofo platónico, sino buscando el equilibrio entre tecnocracia y democracia participativa. Las cuestiones que nos ocupan nos incumben a todos. Por lo tanto, no se pueden dejar en manos de unos pocos, independientemente de si se trata de miembros de un gobierno o de grandes empresas. Esto nos conduce una vez más a preguntarnos cómo hacer que funcione la normativa para la innovación y la participación responsable en la IA. El problema no atañe solo al poder, sino también al bien en general: el bien para con los individuos y para con la sociedad. Quizá nuestras concepciones de una buena vida y una buena sociedad, si es que somos capaces de consensuarlas, requieran una revisión profunda. Permítanme sugerir que, en el caso de Occidente, podría ser útil explorar la idea de aprender de otros sistemas políticos no occidentales y de otras culturas políticas. Una política de actuación respecto de la IA que sea efectiva y esté bien justificada no debería evitar el intervenir en este tipo de discusiones ético-filosóficas y político-filosóficas.

Interdisciplinariedad y transdisciplinariedad

Existen más barreras que necesitamos superar si queremos conseguir una ética de la IA más efectiva y apoyar el desarrollo responsable de la tecnología, evitando lo que los investigadores técnicos llaman un nuevo «invierno» de la IA, es decir, la ralentización del desarrollo y la inversión en este campo. Una de estas barreras es la falta de *interdisciplinariedad* y *transdisciplinariedad.* Aún existe una gran brecha en la formación y comprensión entre la gente de humanidades y ciencias sociales y la gente de ciencias naturales e ingeniería, tanto dentro como fuera del mundo académico. Hasta ahora, el apoyo institucional para tender puentes sustanciales y significativos entre estos dos «mundos» ha sido insuficiente, tanto en el ámbito académico como en la sociedad en general. Pero si realmente queremos tener una tecnología avanzada ética tal como la pretende la ética de la IA, necesitamos acercar ambos mundos, y mejor pronto que tarde.

Para ello es necesario un cambio en nuestras maneras de investigar y de desarrollar (que deberían involucrar no solo a los especialistas técnicos y del mundo de los negocios, sino también a la gente formada en humanidades, por ejemplo), así como en nuestra manera de educar, tanto a los jóvenes como a los no tan jóvenes. Necesitamos asegurarnos de que, por una parte, las personas con una formación en humanidades se vuelvan conscientes de la importancia de reflexionar sobre las nuevas tecnologías como la IA y puedan adquirir algún conocimiento sobre ellas y sus utilidades. Por otro lado, los científicos e ingenieros necesitan ser más sensibles a los aspectos éticos y sociales del desarrollo y uso tecnológico. Cuando aprenden a usar la IA y, más tarde, contribuyen al desarrollo de una nueva tecnología, la ética debería percibirse no como un tema marginal que tiene poco que ver con su práctica tecnológica, sino como *una parte esencial de la misma.* Idealmente, lo que significa «crear IA» o «hacer ciencia de datos» debería simplemente incluir a la ética. De manera más general, podríamos considerar la idea de un tipo más diverso y holista de *Bildung* o de narrativa radicalmente interdisciplinar o pluralista por lo que respecta a los métodos y enfoques, sus temas y, también, sus medios y tecnologías. Por decirlo claramente: si los ingenieros aprendieran a hacer cosas con los textos y la gente de humanidades a hacer cosas con ordenadores, habría más esperanza para una ética y una política de la tecnología que funcionasen en la práctica.

El riesgo de un invierno de la IA y el peligro de un uso excesivo

Si estas indicaciones sobre la política de actuación relativa a la IA y la educación no llegan a llevarse a la práctica y, de forma más general, si el proyecto de una IA ética falla, no nos enfrentamos solo al riesgo de un «invierno de la IA»: el riesgo último y probablemente más importante es un desastre ético, social y económico con sus respectivos costes humanos, no humanos y medioambientales. Esto no tiene tanto que ver con la singularidad, *terminators* u otras especulaciones apocalípticas en torno a un futuro lejano, sino con el lento pero constante aumento de la acumulación de riesgo tecnológico y el resultante crecimiento de vulnerabilidades humanas, sociales, económicas y medioambientales. El incremento en riesgos y vulnera-

bilidades está relacionado con los problemas éticos indicados aquí y en los capítulos previos, incluyendo el uso ignorante y desproporcionado de tecnologías de la automatización como la IA. La brecha en la educación quizás esté agravando los riesgos de la IA, pues, aunque no siempre genere nuevos riesgos, *sí multiplica los riesgos ya existentes.* Por ahora no hay nada parecido a un «carnet de conducir» para utilizar IA, y tampoco una educación obligatoria en ética de la IA para investigadores técnicos, gente de negocios, administradores y personas involucradas en la innovación, la utilización y las políticas de la IA. Hay un montón de IA salvaje en manos de gente que no conoce los riesgos y problemas éticos, o que puede tener unas expectativas equivocadas. El peligro estriba, de nuevo, en el ejercicio del poder sin conocimiento y (por lo tanto) sin responsabilidad: y lo peor es que hay terceras personas expuestas a él. Si existe algo parecido al mal absoluto, habita donde lo situó la filósofa del siglo XX Hannah Arendt: en el sinsentido del trabajo y las decisiones banales cotidianas. Asumir que la IA es neutral y utilizarla sin entender lo que se hace contribuye a ese sinsentido y, en última instancia, a la corrupción ética del mundo. Las políticas educativas pueden ayudar a mitigar esta cuestión y, así, contribuir a una IA buena y significativa.

Sin embargo, aún hay que responder a un número acuciante de preguntas, quizás ligeramente dolorosas, que a menudo se pasan por alto en las discusiones sobre ética y diseño de políticas para la IA, pero que merecen al menos que se las mencione, si no un análisis mucho más profundo. ¿Debe la ética de la IA dedicarse en exclusiva a lo bueno y lo valioso para los seres humanos, o deberíamos, por el contrario, considerar también valores, bienes e intereses no humanos? E, incluso si la ética de la IA debiera dedicarse principalmente a los seres humanos, ¿podría ser que el de la ética de la IA no sea el problema más importante que tenga que abordar la humanidad? Esta cuestión nos lleva al último capítulo.

Capítulo 12

¡Es el clima, imbécil! Sobre prioridades, el Antropoceno y el coche de Elon Musk en el espacio

¿Debería ser antropocéntrica la ética de la IA?

Aunque muchos escritos sobre ética y política de actuación respecto de la IA mencionan el medio ambiente o el desarrollo sostenible, también es cierto que subrayan los valores humanos y son, a menudo, explícitamente antropocéntricos. Por ejemplo, las directrices éticas del HLEG dicen que es necesario un enfoque antropocéntrico de la IA «en el que el ser humano disfrute de un estatus moral de primacía único e inalienable en los campos de lo civil, lo polítco, lo económico y lo social» (Comisión Europea AI HLEG 2019, 10) e, igualmente, universidades como Stanford y el MIT han enmarcado sus políticas de investigación en términos de IA antropocéntrica .

Este antropocentrismo es normalmente definido en relación con la tecnología: la idea es que el bien y la dignidad humana gozan de prioridad sobre cualquier cosa que requiera o haga la tecnología. Lo fundamental es que la tecnología beneficie y sirva a los humanos y no al revés. Sin embargo, como hemos visto en los primeros capítulos, la idoneidad de este enfoque centrado en los humanos no es tan incuestionable como podría parecer en un primer momento, especialmente si tenemos en cuenta los puntos de vista posthumanistas o revisamos críticamente las narrativas competitivas (humanos *vs.* tecnología). La filosofía y la tecnología demuestran que existen otras (a veces más sutiles y sofisticadas) formas de definir la relación entre humanos y tecnología. De igual forma, un enfoque antropocéntrico es, cuando menos, cuestionable, si no problemático, a la luz de los debates filo-

sóficos sobre el medio ambiente y otros seres vivos. En la filosofía y la ética medioambientales existe una prolongada discusión sobre el valor de los seres no humanos, en especial los vivos, en torno a cómo respetar dicho valor y dichos seres, y sobre las tensiones potenciales que pudieran surgir en relación con los valores humanos. Para la ética de la IA, esto implica que deberíamos formular la pregunta sobre el impacto de la IA en otros seres vivos y considerar el problema de que pueda haber una tensión entre valores e intereses humanos y no humanos.

Entendiendo nuestras prioridades

También podría argumentarse que existen problemas más graves que los causados por la IA, y que es importante entender correctamente nuestras prioridades. Esta objeción podría surgir al pensar en problemas globales como el cambio climático, el principal problema, en opinión de muchos, que la humanidad tiene que abordar y priorizar dada su urgencia y potencial impacto en el conjunto del planeta.

Viendo la agenda de las Naciones Unidas de 2015 para el desarrollo sostenible (los llamados «Objetivos de desarrollo sostenible») y su aproximación general a los problemas globales relativos a lo que el secretario general de las Naciones Unidas Ban Ki-moon llamó «personas y planeta», observamos numerosos problemas que requieren de atención ética y política urgente: el aumento de las desigualdades dentro y entre países, la guerra y el extremismo violento, la pobreza y la malnutrición, la falta de acceso a agua potable, la falta de instituciones efectivas y democráticas, el envejecimiento de las poblaciones, las infecciones y las enfermedades epidémicas, los riesgos relacionados con la energía nuclear, la falta de oportunidades para los niños y la gente joven, las desigualdades de género y las distintas formas de discriminación y exclusión, las crisis humanitarias y todo tipo de violaciones de los derechos humanos, conflictos relacionados con la migración y los refugiados, y el cambio climático y los problemas medioambientales (a menudo relacionados con el cambio climático), como los cada vez más frecuentes e intensos desastres naturales y otras formas de degradación medioambiental como sequías y pérdida de biodiversidad. A la luz de estos

enormes problemas, ¿debería ser la IA nuestra primera prioridad? ¿Nos aparta la IA de problemas más importantes?

Un enfoque antropocéntrico es, cuando menos, cuestionable, si no problemático, a la luz de los debates filosóficos sobre el medio ambiente y otros seres vivos.

Por una parte, concentrarse en la IA y en otros problemas relacionados con la tecnología parece estar fuera de lugar cuando tantas personas sufren y tantos problemas afectan al mundo. Mientras la gente de una parte del mundo lucha por conseguir el acceso al agua potable o por sobrevivir en entornos violentos, la de la otra se preocupa por su privacidad en internet y fantasea con un futuro en el que la IA alcance la superinteligencia. En términos éticos, parece que algo no cuadra, algo relacionado con las desigualdades globales y las injusticias. La ética y la política no deberían ser ciegas a dichos problemas, que no tienen por qué tener que ver con la IA. Por ejemplo, a veces en países en vías de desarrollo una tecnología básica puede ser de más utilidad para lidiar con los problemas de las personas que una de orden superior, ya que se pueden permitir instalarla y mantenerla.

Por otra parte, la IA podría causar nuevos problemas y también *agravar los ya existentes* en las sociedades y el medio ambiente. Por ejemplo, algunos temen que la IA ensanche la brecha entre ricos y pobres y que, como muchas tecnologías digitales, incremente el consumo de energía y origine más residuos. Desde esta perspectiva, discutir y lidiar con los problemas éticos de la IA no constituye una distracción, sino una de las formas de las que disponemos para hacer frente a los problemas del mundo, incluyendo los medioambientales. Se podría así concluir que *también* necesitamos prestar atención a la IA: sí, la pobreza, la guerra, *etc.* son problemas graves, pero la IA puede también causar o agravar conflictos serios ahora y en el futuro, y debería estar en nuestra lista de problemas que requieren solución. Sin embargo, esto no responde a la pregunta sobre las prioridades (una pregunta ética y políticamente importante). Y el problema no está en que haya o no respuestas fáciles, sino en que la mayoría de escritos académicos y documentos dedicados a las políticas de actuación con respecto a la IA no se han hecho esta pregunta. La cuestión es que ni siquiera se ha *hecho la pregunta* en la mayoría de escritos académicos y en los documentos que tratan sobre las políticas a llevar a cabo respecto de la IA.

Mientras la gente de una parte del mundo lucha por conseguir acceso a agua potable o por sobrevivir en entornos violentos, la de la otra se preocupa por su privacidad en internet.

IA, CAMBIO CLIMÁTICO Y ANTROPOCENO

Una de las formas más estimulantes de hacernos la pregunta sobre las prioridades es introducir en la discusión el cambio climático u otros temas relacionados, como el del Antropoceno: «¿Por qué preocuparnos por la IA cuando el problema más urgente es el del cambio climático y está en juego el futuro del planeta?»; o, mejor, y adaptando una expresión de cultura política política estadounidense: «¡Es el clima, imbécil!». Examinemos la cuestión y discutamos qué implica para la reflexión sobre la ética de la IA.

Aunque algunos extremistas rechazan las evidencias científicas, el cambio climático está ampliamente reconocido por los científicos y los diseñadores de políticas de actuación, no solo como un problema global serio, sino como «uno de los grandes desafíos de nuestro tiempo», como se señala en los objetivos de desarrollo sostenible de las Naciones Unidas. No es solo un problema del futuro: la temperatura global y los niveles del mar *ya* están subiendo, y está afectando a las zonas y países costeros con menor altitud. Pronto cada vez más gente tendrá que enfrentarse a las consecuencias del cambio climático. La conclusión de muchos es que tenemos que actuar urgentemente, *ahora,* para mitigar los efectos del cambio climático: «mitigar» porque el proceso puede que ya haya pasado el punto de inflexión. La idea es que no solo ya es hora de hacer algo, sino que posiblemente ya sea demasiado tarde para evitar todas las consecuencias. Comparada con los miedos del transhumanismo a la superinteligencia, esta preocupación está más respaldada por evidencias científicas y ha ganado un apoyo considerable entre las élites cultas de Occidente, que, comprensiblemente aburridas del escepticismo postmoderno y las políticas burocratizadas de la identidad, ven ahora una razón para centrarse en un problema que parece tan cierto, tan real y tan universal: el cambio climático está ocurriendo *realmente* y afecta a *todo el mundo* y a todo en el planeta. Una reciente ola de activismo llama la atención de la crisis climática, la campaña de Greta Thunberg y las huelgas climáticas son ejemplos de ello.

A veces se usa el concepto de «Antropoceno» para señalar este problema. Acuñado por el investigador del clima Paul Crutzen y el biólogo Eugene Stoermer, parte de la idea de que estamos viviendo en una época geológica en la que la humanidad ha incrementado drásticamente su poder con respecto a la Tierra y sus ecosistemas, convirtiendo a los seres humanos en una fuerza geológica. Hay que tener en cuenta, en este sentido, el crecimiento exponencial de las poblaciones tanto humanas como del ganado, el uso masivo de agua potable, la extinción de especies, la emisión de sustancias tóxicas, *etc.* Algunos piensan que el Antropoceno comenzó con la revolución agrícola; otros que lo hizo con la revolución industrial (Crutzen 2006) o tras la Segunda Guerra Mundial. En cualquier caso, se han creado un nuevo relato y una nueva historia, quizás incluso una nueva gran narrativa. El concepto se utiliza a menudo para generar inquietud en torno al calentamiento global y el cambio climático, y para impulsar en distintas disciplinas (incluyendo las humanidades) la reflexión sobre el futuro del planeta.

¿Por qué preocuparnos por la IA cuando el problema más urgente es el del cambio climático y está en juego el futuro del planeta?

No todo el mundo adopta este término (es controvertido incluso entre los geólogos) y algunos han cuestionado su antropocentrismo. Por ejemplo, Haraway (2015) defendió desde una perspectiva posthumanista que otras especies y actores «abióticos» también desempeñan un papel en el cambio del entorno. Pero incluso sin un concepto controvertido como el del Antropoceno, el cambio climático y otros problemas medioambientales han llegado para quedarse, y la política debe lidiar con ellos, mejor pronto que tarde. ¿Qué implica esto para la política referente a la IA?

Numerosos investigadores creen que la IA y el *big data* también podrían ayudarnos a lidiar con los problemas globales, incluido el cambio climático. Como sucede con las tecnologías de la información digital y de la comunicación en general, la IA puede contribuir a un desarrollo sostenible y a afrontar muchos problemas medioambientales. Es muy probable, de hecho, que la de la IA sostenible se convierta en una vía exitosa de investigación y desarrollo. Sin embargo, la IA también podría empeorar las cosas para el medio ambiente y, por tanto, para todos. Considérese de nuevo el aumento del consumo de energía y de residuos. Además, desde la perspectiva de los defensores del Antropoceno, el riesgo estriba en que los humanos puedan usar la IA para fortalecer su control sobre la Tierra, empeorando, así, el problema en lugar de solucionarlo.

Esto resulta especialmente problemático cuando la IA es vista no solo como *una* solución, sino como la solución *principal*. Considérese el escenario de la superinteligencia en el que una IA sabe mejor que nosotros lo que es bueno para los humanos: una IA «benigna» que sirve a la humanidad haciendo que los individuos actúen de acuerdo con sus intereses y con los del planeta; es decir, un equivalente tecnológico al rey-filósofo de Platón, un dios máquina. Un *homo deus* (Harari 2015) es sustituido por un *IA deus,* que controle tanto nuestro sistema de apoyo vital como a nosotros mismos. Para resolver los problemas de la distribución de recursos, por ejemplo, la IA podría actuar como una especie de «servidor», controlando el acceso que tienen los humanos a los recursos. Sus decisiones estarían basadas en sus análisis de patrones de datos. Este régimen podría combinarse con solucio-

nes tecnológicas prometeicas, como la geo-ingeniería. Los seres humanos no son los únicos que necesitan un control, también el planeta debe ser rediseñado. La tecnología se emplearía, de este modo, para «arreglar» nuestros problemas y los del planeta.

Sin embargo, estos escenarios no solo serían autoritarios y violarían la autonomía humana, sino que también contribuirían principalmente al problema del Antropoceno mismo: la hiperagencia humana, esta vez delegada por los humanos en las máquinas, convertiría al planeta entero en un recurso y una máquina para los humanos. El problema del Antropoceno se «resuelve» llevándolo a su extremo tecnocrático, conduciendo a un mundo de máquinas en el que los humanos son tratados primero como niños a los que hay que cuidar y que quizás se acaben quedando obsoletos. Con este tipo de Antropoceno basado en el *big data* y el demasiado familiar drama de los humanos sustituidos por máquinas, volvemos a los escenarios de sueños y pesadillas.

La nueva locura espacial y la tentación platónica

Otra respuesta al cambio climático y el Antropoceno, que también pasa por una visión tecnófila y a menudo ligada a las narrativas transhumanistas, es la siguiente: si destruimos este planeta, podríamos escapar de la Tierra e irnos al espacio.

Una imagen icónica de 2018 fue la del coche deportivo marca Tesla de Elon Musk flotando en el espacio . Musk también tiene planes para colonizar Marte. No es el único soñador: hay un creciente interés en ir al espacio. Y es más que un simple sueño: se está invirtiendo mucho dinero en programas espaciales. En contraste con la carrera espacial del siglo XX, esta está impulsada por empresas privadas. Y no solo están muy interesados en el espacio los millonarios de la tecnología sino también los artistas. La empresa de Musk SpaceX planea enviar artistas a dar vueltas alrededor de la órbita de la luna . El turismo espacial es otra idea cada vez más popular. ¿Por qué no desear ir al espacio? Al fin y al cabo, es atractivo.

No hay nada malo con ir al espacio *per se,* y ciertamente tiene beneficios potenciales. Por ejemplo, investigar modos de supervivencia en entornos

extremos nos puede ayudar a enfrentarnos con problemas que tenemos en la Tierra, e igualmente son útiles los experimentos con energías sostenibles, así como adoptar una perspectiva planetaria. Considérese también que el problema del Antropoceno pudo formularse solo porque, muchos años antes, la tecnología espacial nos hizo ver la Tierra desde una cierta distancia. Y volviendo de nuevo al coche de Musk, lo cierto es que algunos piensan que el coche eléctrico es una solución a los problemas ambientales, sin cuestionar la suposición de que los coches son el mejor medio de transporte y sin tener en cuenta cómo se produce la electricidad. En cualquier caso, se están proponiendo ideas interesantes.

Pero los sueños espaciales resultan problemáticos si lo que se consigue es que se descuiden los problemas terrestres, y si son sintomáticos de una tendencia que ya diagnosticó Hannah Arendt en 1958, cuando escribió, refiriéndose a la condición humana, que sufre de demasiada abstracción y alienación. Sugirió que la ciencia apoya el deseo de dejar la Tierra: literalmente, por medio de la tecnología espacial (en su tiempo, el Sputnik), pero también mediante métodos matemáticos que abstraen y alienan de lo que yo llamaría nuestra vida meramente terrestre, encarnada y política. Desde esta perspectiva, las fantasías transhumanistas sobre la superinteligencia y el abandono de la Tierra pueden ser interpretadas como exponentes de una forma problemática de alienación y escapismo. Es el platonismo y el transhumanismo con mayúsculas: la idea es superar no solo las limitaciones del cuerpo humano, sino también las de ese otro «sistema de soporte vital»: la Tierra misma. No solo el cuerpo, sino también la Tierra, son vistos como una prisión de la que hay que escapar.

Un peligro de la IA, desde este punto de vista, es que facilita este tipo de pensamiento y se convierte en una máquina de alienación: un instrumento para abandonar la Tierra y rechazar nuestra condición vulnerable, corpórea, terrestre y existencialmente dependiente. En otras palabras: un cohete. De nuevo, no hay nada malo en los cohetes *per se.* El problema es la mezcla de tecnologías concretas con narrativas concretas. Aunque la IA puede suponer una fuerza potencialmente positiva para nuestras vidas personales, la sociedad y la humanidad, una combinación de las tendencias abstraccionistas y alienantes de la ciencia y la tecnología con las fantasías transhumanistas y

«trans-tierristas» puede conducir a un futuro tecnológico negativo para los seres humanos y otros seres vivos de la Tierra. Si escapamos en vez de enfrentarnos a nuestros problemas (por ejemplo, al cambio climático), puede que conquistemos Marte (por ahora), pero perdamos la Tierra.

Además, y como siempre, este asunto también tiene una lectura política: algunos tienen más oportunidades, dinero y poder que otros para escapar. El problema no es solo que la tecnología espacial y la IA tengan costes reales para la Tierra y que el dinero invertido en los proyectos espaciales no se invierta en problemas reales terrestres como la guerra o la pobreza: la preocupación estriba también en que los ricos serían capaces de escapar de la Tierra que destruyen, mientras que el resto de nosotros tendría que permanecer en un planeta cada vez más inhabitable (véase, por ejemplo, Zimmerman 2015). Como los cohetes y otras tecnologías, la IA puede convertirse en una herramienta para la «supervivencia de los más ricos», como expresó Rushkoff (2018). De hecho ya sucede algo parecido a esto con otras tecnologías: en ciudades como Delhi y Beijing, la mayoría de la gente está rodeada de contaminación atmosférica mientras los ricos se libran por vivir en zonas menos contaminadas, o compran aire de calidad utilizando tecnologías de purificación del aire. No todo el mundo respira lo mismo. ¿Contribuirá la IA a mantener las brechas entre ricos y pobres, propiciando que la vida sea más estresante e insalubre para unos y mejor para otros? ¿Nos alienará la IA de los problemas medioambientales? Parece exigible éticamente que la IA mejore la vida en la Tierra, preferiblemente para todos y teniendo en cuenta que la existencia humana depende de ella. Algunas narrativas espaciales pueden obstaculizarnos, más que ayudarnos, a alcanzar esta meta.

Vuelta a la Tierra: hacia una IA sostenible

Pero volvamos al problema práctico de las prioridades y a los riesgos, muy reales e inmediatos, del cambio climático. ¿Qué caminos deberían seguir la ética y las políticas de actuación respecto de la IA a la luz de estos desafíos? Y, cuando hay conflictos con el valor de las vidas no humanas, ¿cómo deben resolverse? La mayoría de la gente estará de acuerdo en que entregar el control a una IA o escapar de la Tierra no son buenas soluciones.

¿Pero qué es una buena solución? ¿Hay una solución acaso? Una respuesta productiva a estas cuestiones nos lleva de vuelta a las preguntas filosóficas relacionadas con el modo por el cual nosotros, como seres humanos, nos relacionamos con la tecnología y con nuestro entorno. También nos lleva de vuelta al capítulo de la tecnología: ¿qué pueden hacer la IA y la ciencia de datos por nosotros, y qué podemos esperar razonablemente de la IA?

Está claro que la IA puede ayudarnos a combatir los problemas medioambientales. Considérese el cambio climático. La IA podría parecer especialmente adecuada para ayudarnos con problemas tan complejos, pues nos puede ayudar a estudiar el problema, por ejemplo, al detectar patrones en los datos medioambientales imposibles de apreciar para nosotros por la abundancia de los datos y su complejidad. También puede servirnos para encontrar soluciones, por ejemplo, ayudándonos a lidiar con la organización de la complejidad y a implementar medidas para la reducción de emisiones nocivas, como han aducido Floridi *et al.* (2018). Más generalmente, una IA podría servir de ayuda monitorizando y modelando sistemas ambientales y facilitando soluciones tales como redes de energía y agricultura inteligentes, como propone un blog del World Economic Forum (Herweijwer 2018). Los gobiernos, pero también las empresas, deberían asumir el mando en este sentido. Por poner un caso, Google ya utiliza la IA para reducir el uso de energía en centros de datos.

Sin embargo, esto no tiene por qué significar «la salvación del planeta». La IA también puede generar problemas y empeorar las cosas. Considérese de nuevo el impacto medioambiental negativo que puede tener si la energía, la infraestructura y los materiales dependen de ella. Necesitamos valorar no solo el uso sino también la producción: la electricidad puede producirse de formas poco o nada sostenibles, y la producción de dispositivos de IA consume energía, materias primas y genera residuos. También puede pensarse, en este sentido, en el «auto-empujón» propuesto por Floridi et al: sugieren que la IA puede ayudarnos a comportarnos de formas medioambientalmente correctas contribuyendo a que nos mantengamos firmes en nuestras decisiones. Pero esto conlleva sus propios riesgos éticos, pues no está claro que respete la autonomía y la dignidad humanas; como afirman los autores, cabría la posibilidad de que se situase en la misma dirección que aquella IA

benigna, de la que ya hemos hablado, que cuida de los seres humanos pero destruye su libertad y contribuye al problema del Antropoceno. Cuando menos, existe el *riesgo* de nuevas formas de paternalismo y autoritarismo. Además, usar la IA para enfrentarse al cambio climático puede conllevar una visión del mundo en la cual este quede reducido a un mero repositorio de datos, así como una de los humanos que reduzca su inteligencia al procesamiento de datos (quizás, incluso, a un tipo de procesamiento de inferior, que necesitaría mejorarse mediante la intervención de máquinas). Es improbable que tales visiones reformulen nuestra relación con el entorno de una forma que mitigue el problema que suponen hechos como el cambio climático y el resto de los que se engloban en la idea del Antropoceno.

También nos enfrentamos al riesgo del tecno-solucionismo, en el sentido de que las propuestas para usar la IA para solucionar los problemas medioambientales tienden a suponer que existe una solución final a todos los problemas, que la tecnología por sí sola puede dar respuesta a las preguntas más difíciles y que es posible resolver los problemas utilizando solo la inteligencia humana o artificial. Pero los problemas medioambientales no se pueden resolver únicamente mediante la inteligencia científico-tecnológica: también están ligados a problemas políticos y sociales complejos. Los problemas medioambientales son siempre problemas humanos. Y aunque la matemática y su vástago tecnológico son herramientas muy útiles, están limitadas a la hora de comprender los problemas humanos y enfrentarse a ellos. Por ejemplo, los valores pueden entrar en conflicto. La IA no nos ayuda necesariamente a resolver la cuestión de las prioridades, que es una cuestión de índole ética y política importante y exige una respuesta humana. Además, las humanidades y las ciencias sociales nos advierten que hay que ser muy cuidadosos con las soluciones «finales».

Además, los seres humanos no son los únicos que tienen problemas: los no humanos también se enfrentan a dificultades, que a menudo son desatendidas en las discusiones sobre el futuro de la IA. Finalmente, la idea de que deberíamos escapar de la Tierra, o la de un mundo en el cual todo consiste en datos que los humanos podemos manipular con la ayuda de máquinas, puede conducir a ampliar la brecha entre ricos y pobres y a formas de explotación y violación a gran escala de la dignidad humana, así como amena-

zar las vidas de generaciones futuras por la destrucción de las condiciones de vida en nuestro planeta. Debemos reflexionar más en profundidad sobre cómo construir sociedades y entornos sostenibles: necesitamos pensamiento *humano.*

Se buscan: inteligencia y sabiduría

Con todo, el modo en que pensamos los seres humanos tiene también varias caras. La IA está relacionada con un tipo de pensamiento y de inteligencia humanos: el más abstracto, el de tipo cognitivo. Dicho tipo de pensamiento ha resultado muy exitoso, pero tiene limitaciones y no es el único que podemos o debemos valorar. Para responder a cuestiones éticas y políticas sobre cómo vivir, cómo lidiar con nuestro entorno y cómo relacionarnos mejor con seres no humanos se necesita algo más que la inteligencia humana abstracta (por ejemplo, argumentos, teorías, modelos) o el reconocimiento de patrones de la IA. Necesitamos personas ingeniosas y máquinas inteligentes, pero también intuiciones y conocimientos prácticos que no pueden explicitarse del todo, así como un saber pragmático y depurado capaz de responder ante situaciones y problemas concretos a fin de determinar nuestras prioridades. Tal saber podría conformarse a partir de procesos cognitivos abstractos y de análisis de datos, pero también a partir de experiencias corporeizadas, relacionales y situacionales, de lidiar con otras personas, con la materialidad y con entornos naturales distintos. Nuestro éxito al enfrentarnos con los grandes problemas de nuestro tiempo dependerá seguramente de una combinación de inteligencia abstracta (humana y artificial) y sabiduría pragmática concreta desarrollada sobre la base de experiencia humana situacional y práctica (incluyendo nuestra experiencia con la tecnología). Sea cual sea la dirección que tome el desarrollo subsecuente de la IA, el desafío de desarrollar este último tipo de conocimiento y aprendizaje es nuestro. Los humanos tenemos que hacerlo. La IA es buena a la hora de reconocer patrones, pero la sabiduría no se puede delegar en las máquinas.

Glosario

Agencia moral: capacidad para la acción, razonamiento, juicio y toma de decisiones morales, opuesta a una simple consecuencia moral.

Antropoceno: presunta época geológica actual en la que la humanidad ha incrementado drásticamente su poder y efecto sobre la Tierra y sus ecosistemas, convirtiéndose a sí misma en una fuerza geológica.

Aprendizaje automático: máquina o *software* capaz de aprender automáticamente, no en la forma en que aprendemos los humanos, sino a partir de un proceso computacional y estadístico. Alimentándose de datos, los algoritmos de aprendizaje pueden detectar patrones o reglas en los datos y hacer predicciones para datos futuros.

Aprendizaje profundo: forma de *aprendizaje automático* que utiliza redes neuronales con varias capas de «neuronas» (unidades interconectadas de procesamiento simple que interactúan entre sí).

Ciencia de datos: ciencia interdisciplinar que utiliza la estadística, algoritmos y otros métodos para extraer patrones significativos y útiles de conjuntos de datos (a veces conocidos como *big data*). Hoy en día, el *aprendizaje automático* se usa a menudo en este campo. Junto con el análisis de datos, la ciencia de datos también se ocupa de recolectar, preparar e interpretar los datos.

Ética en virtud del diseño: un enfoque de la ética de la tecnología y un componente clave de la *innovación responsable* que busca integrar la ética en la fase de diseño y desarrollo de la tecnología. A veces también se hace referencia a este concepto mediante la expresión «valores integrados en el diseño». Son conceptos similares «diseño sensible a los valores» y «diseño éticamente alineado».

Ética positiva: ética que se preocupa por cómo debemos vivir (juntos), a partir de una visión buena de la vida y de la sociedad. Contrasta con la ética negativa, que establece límites e indica lo que no deberíamos hacer.

Explicabilidad: la capacidad de explicar o de ser explicado. En el contexto de la ética, se refiere a la capacidad de explicar a otros por qué se ha hecho algo o por qué se ha tomado una decisión; es parte de lo que implica ser responsable.

Inteligencia artificial (IA): inteligencia demostrada o simulada mediante medios tecnológicos. A menudo se considera «inteligente» a aquello que responde a ciertos estándares de la inteligencia humana, es decir, al tipo de capacidades inteligentes y comportamientos que presentan los seres humanos. El término puede referirse a la ciencia o a la tecnología, como sucede, por ejemplo, en el caso de los algoritmos de aprendizaje.

IA fiable: IA en la que pueden confiar los humanos. Las condiciones para tal confianza se pueden referir a (otros) principios éticos tales como la dignidad humana, el respeto de los derechos humanos, etc., y/o a los factores sociales y técnicos que influyen en si las personas quieren o no utilizar la tecnología. El uso del término «fiable» en relación con las tecnologías es controvertido.

IA explicable: IA capaz de explicar a los humanos sus acciones, decisiones o recomendaciones, o de facilitar información suficiente sobre los procesos por los cuales llegó a un resultado en particular.

IA general: inteligencia similar a la humana, que se puede aplicar ampliamente y en oposición a la IA «estrecha», que solo podría aplicarse a un problema o tarea concretos. También se denomina IA «fuerte» en oposición a la IA «débil».

IA simbólica: IA que se basa en la representación simbólica de tareas cognitivas superiores como el razonamiento abstracto y la toma de decisiones. Puede utilizar un árbol de decisión y adoptar la forma de un sistema especializado que necesita *inputs* de expertos en el campo en cuestión.

IA sostenible: IA que permite y contribuye a una forma de vida sostenible para la humanidad y que no destruye los ecosistemas de la Tierra, de los que dependen los seres humanos (y también muchos no humanos).

Innovación responsable: enfoque que persigue una innovación más responsable ética y socialmente, que en general pasa por integrar la ética en el diseño y por tomar en consideración las opiniones e intereses de las partes interesadas.

Paciencia moral: el estatus moral de una entidad en el sentido de cómo debería tratarse esa entidad.

Posthumanismo: Un abanico de creencias que cuestionan el humanismo, especialmente la posición central del ser humano, y que expande el círculo de la preocupación ética a los seres no humanos.

Responsabilidad moral: puede usarse como sinónimo de «ser moral» y se referirá, entonces, a producir consecuencias moralmente buenas, adherirse a principios morales, ser virtuoso, ser merecedor de elogio, *etc.* (el énfasis depende de la teoría normativa aceptada). Se puede también preguntar en qué condiciones puede alguien ser considerado responsable. Las condiciones para atribuir responsabilidad moral son la *agencia moral* y el conocimiento. Los enfoques relacionales subrayan que siempre se rinden cuentas ante otros.

Sesgo: discriminación en contra o a favor de individuos o grupos concretos. En el contexto de la ética y de la política, la cuestión que se plantea gira en torno a si un sesgo determinado es injusto o no equitativo.

Singularidad tecnológica: la idea de que llegará un momento en la historia de la humanidad en el que se producirá una explosión en la inteligencia de las máquinas que conllevará cambios tan drásticos para nuestra civilización que dejaremos de comprender lo que ocurre.

Superinteligencia: la idea de que las máquinas superarán la inteligencia humana; a veces está conectada con la idea de una «explosión de inteligencia» vinculada a máquinas inteligentes que diseñan máquinas cada vez más inteligentes.

Transhumanismo: la creencia de que los humanos deberían mejorarse a sí mismos mediante tecnologías avanzadas y, de esta forma, transformar la condición humana: la humanidad debería pasar a la siguiente fase. Es también un movimiento internacional.

Bibliografía

ACCESSNOW, «Mapping Regulatory Proposals for Artificial Intelligence in Europe», 2018. Disponible en <https://www.accessnow.org/cms/assets/uploads/2018/11/mapping_regulatory_proposals_for_AI_in_EU.pdf>.

ACRAI (Austria Council on Robotics and Artificial Intelligence), «Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positive gestalten: White paper des Österreichischen Rats für Robotik und Künstliche Intelligenz», 2018.

«Algorithm and Blues», *Nature* 537: 449, 2016.

ALGORITHM WATCH, «Automating Society: Taking Stock of Automated Decision Making in the EU» (informe de Algorithm Watch en cooperación con Bertelsmann Stiftung), enero 2019, Berlín, AW GmbH, 2019. Disponible en <http://www.algorithmwatch.org/automating-society>.

ALPAYDIN, Ethem, *Machine Learning,* Cambridge, MA, MIT Press, 2016.

ANDERSON, Michael y ANDERSON, Susan, «General Introduction», *Machine Ethics,* editado por Michael Anderson y Susan Anderson, Cambridge, Cambridge University Press, 2011, 1-4.

ARENDT, Hannah, *La condición humana,* Barcelona, Paidós, 1993.

ARKOUDAS, Konstantine y BRINGSJORD, Selmer, «Philosophical Foundations», *The Cambridge Handbook of Artificial Intelligence,* editado por Keith Frankish y William M. Ramsey, Cambridge, Cambridge University Press, 2014.

ARMSTRONG, Stuart, *Smarter Than Us: The Rise of Machine Intelligence,* Berkeley, Machine Intelligence Research Institute, 2014.

AWAD, Edmond; DSOUZA, Sohan; KIM, Richard; SCHULZ, Jonathan; HENRICH, Joseph; SHARIFF, Azim; BONNEFON, Jean-François y RAHWAN, Iyad, «The Moral Machine Experiment». *Nature* 563:59-64, 2018.

BACON, Francis, *Refutación de las filosofías,* Madrid, Encuentro, 2014.

BODDINGTON, Paula, «The Distinctiveness of AI Ethics, and Implications for Ethical Codes» (trabajo presentado en el taller Ética para la inteligencia artificial (Ethics for Artificial Intelligence), 9 de julio de 2016, IJCAI- 16, Nueva York, 2016. Disponible en <https://www.cs.ox.ac.uk/efai/2016/11/02/the-distinctiveness-of-ai-ethics-and-implications-for-ethical-codes/>.

BODDINGTON, Paula, *Towards a Code of Ethics for Artificial Intelligence,* Cham, Springer, 2017.

BODEN, Margaret A., *AI: Its Nature and Future,* Oxford, Oxford University Press, 2016.

BOROWIEC, Steven, «AlphaGo Seals 4-1 Victory Over Go Grandmaster Lee Sedol», *Guardian,* 15 de marzo de 2016. Disponible en <https://www.the guardian.com/technology/2016/mar/15/googles-alphago-seals-4-1-victoryover-grandmaster-lee-sedol>.

BOSTROM, Nick, *Superintelligence,* Oxford, Oxford University Press, 2014.

BRYNJOLFSSON, Erik, y MCAFEE, Andrew, *The Second Machine Age,* Nueva York, W. W. Norton, 2014.

BRYSON, Joanna, «Robots Should Be Slaves», *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues,* editado por Yorick Wilks, Amsterdam, John

Benjamins, 2010, 63-74.
—, «AI & Global Governance: No One Should Trust AI», United Nations University Centre for Policy Research, *AI & Global Governance,* 13 de noviembre de 2018. Disponible en <https://cpr.unu.edu/ai-global-governance-no-one-should-trust-ai.html>.
BRYSON, Joanna; DIAMANTIS, Mihailis E. y GRANT, Thomas D., «Of, For, and By the People: The Legal Lacuna of Synthetic Persons», *Artificial Intelligence & Law* 25, núm. 3, 2017, 273-291.
CALISKAN, Aylin; BRYSON, Joanna J. y NARAYANAN, Arvind, «Semantics Derived Automatically from Language Corpora Contain Human-like Biases», *Science* 356, 2017, 183-186.
CASTELVECCHI, Davide, «Can We Open the Black Box of AI?», *Nature* 538, núm. 7623: 2016, 21-23.
CDT (Centre for Democracy & Technology), «Digital Decisions», 2018. Disponible en <https://cdt.org/issue/privacy-data/digital-decisions/>.
COECKELBERGH, Mark, «Moral Appearances: Emotions, Robots, and Human Morality», *Ethics and Information Technology* 12, núm. 3, 2010, 235-241.
COECKELBERGH, Mark, «You, Robot: On the Linguistic Construction of Artificial Others», *AI & Society* 26, núm. 1, 2011, 61-69.
—, *Growing Moral Relations: Critique of Moral Status Ascription,* Nueva York, Palgrave Macmillan, 2012.
—, *Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations,* Cham, Springer, 2013.
—, *New Romantic Cyborgs,* Cambridge, MA, MIT Press, 2017.
CONSEJO DE ESTADO DE CHINA, «New Generation Artificial Intelligence Development Plan». Plan» (trad. al inglés de Flora Sapio, Weiming Chen y Adrian Lo), 2017. Disponible en <https://flia.org/notice-state-council-issuing-new-generation-artificial-intelligence-development-plan/>.
CRAWFORD, Kate y CALO, Ryan, «There Is a Blind Spot in AI Research», *Nature* 538, 2016, 311-313.
CRUTZEN, Paul J., «The "Anthropocene"», *Earth System Science in the Anthropocene,* editado por Eckart Ehlers y Thomas Krafft, 13-18, Cham, Springer, 2006.
DARLING, Kate; NANDY, Palash y BREAZEAL, Cynthia, «Empathic Concern and the Effect of Stories in Human-Robot Interaction», *24th IEEE International Symposium on Robot and Human Interactive Communication (RO- MAN)*, Nueva York, IEEE, 2015, 770-775.
DENNETT, Daniel C., «Consciousness in Human and Robot Minds», *Cognition, Computation, and Consciousness*, editado por Masao Ito, Yasushi Miyashita, y Edmund T. Rolls, Nueva York, Oxford University Press, 1997, 17-29.
DIGITAL EUROPE, «Recommendations on AI Policy: Towards a Sustainable and Innovation-friendly Approach», *Digitaleurope.org,* 7 de noviembre de 2018.
DIGNUM, Virginia; BALDONI, Matteo; BAROGLIO, Cristina; CAON, Maruiyio; CHATILA, Raja; DENNIS, Louise y GÉNOVA, Gonzalo, et al., «Ethics by Design: Necessity or Curse?», Association for the Advancement of Artificial Intelligence, 2018. Disponible en <http://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_68.pdf>.
DIRECCIÓN GENERAL DE EMPLEO, ASUNTOS SOCIALES E INCLUSIÓN DE LA COMISIÓN EUROPEA, «Employment and Social Developments in Europe 2018», Luxemburgo, Publications Office of the European Union, 2018. Disponible en <http://ec.europa.eu/social/main.jsp?catId=738&langId= en&pubId=8110>.
DOWD, Maureen, «Elon Musk's <Billion-Dollar Crusade to Stop the A.I. Apocalypse», *Vanity Fair,* 26 de marzo de 2017. Disponible en <https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x>.

DREYFUS, Hubert L., *What Computers Can't Do,* Nueva York, Harper Collins, 1972.
DRUGA, Stefania y WILLIAMS, Randi, «Kids, AI Devices, and Intelligent Toys», MIT Media Lab, 6 de junio de 2017. Disponible en <https://www.media.mit.edu/posts/kids-ai-devices/f>.
EUROPEAN COMMISSION, «Ethics and Data Protection», 2018. Disponible en <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf>.
EUROPEAN COMMISSION AI HLEG (High-Level Expert Group on Artificial Intelligence), «Draft Ethics Guidelines for Trustworthy AI: Working Document for Stakeholders», Bruselas, European Commission, 2018. Disponible en <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai>.
—, «Ethics Guidelines for Trustworthy AI», Bruselas, European Commission, 8 de abril de 2019. Disponible en <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>.
EGE (European Group on Ethics in Science and New Technologies), «Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems», Bruselas, European Commission, 2018.
EXECUTIVE OFFICE OF THE PRESIDENT, NATIONAL SCIENCE AND TECHNOLOGY COUNCIL COMMITTEE ON TECHNOLOGY, «Preparing for the Future of Artificial Intelligence». Washington, DC, Office of Science and Technology Policy (OSTP), 2016.
FLORIDI, Luciano; COWLS, Josh; BELTRAMETTI, Monica; CHATILA, Raja; CHAZERAND, Patrice; DIGNUM, Virginia; LUETGE, Christoph; MADELIN, Robert; PAGALLO, Ugo; ROSSI, Francesca; SCHAFER, Burkhard; VALCKE, Peggy y VAYENA, Effy, «AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations», *Minds and Machines* 28, núm. 4, 2018, 689-707.
FLORIDI, Luciano y SANDERS, J. W., «On the Morality of Artificial Agents», *Minds and Machines* 14, núm. 3: 2004, 349-379.
FORD, Martin, *Rise of the Robots: Technology and the Threat of a Jobless Future,* Nueva York, Basic Books, 2015.
FRANKISH, Keith y RAMSEY, William M., «Introduction», *The Cambridge Handbook of Artificial Intelligence,* editado por Keith Frankish y William M. Ramsey, Cambridge, Cambridge University Press, 2014, 1-14.
FREY, Carl Benedikt y OSBORNE, Michael A., «The Future of Employment: How Susceptible Are Jobs to Computerisation?», Oxford, Oxford Martin Programme on Technology and Employment (University of Oxford), 2013.
FRY, Hannah, *Hola mundo: cómo seguir siendo humanos en la era de los algoritmos,* Barcelona, Blackie Books, 2019.
FUCHS, Christian, *Digital Labour and Karl Marx,* Nueva York, Routledge, 2014.
GOEBEL, Randy; CHANDER, Ajay; HOLZINGER, Katharina; LECUE, Freddy; AKATA, Zeynep; STUMPF, Simone; KIESEBERG, Peter y HOLZINGER, Andreas, 2018. «Explainable AI: The New 42?» Paper presentado en el CD- MAKE, Hamburg, Germany, agosto de 2018.
GUNKEL, David, *The Machine Question,* Cambridge, MA, MIT Press, 2012.
—, «The Other Question: Can and Should Robots Have Rights?» *Ethics and Information Technology* 20, 2018, 87-99.
HARARI, Yuval Noah, *Homo Deus: A Brief History of Tomorrow,* Londres, Hervill Secker, 2015.
HARAWAY, Donna, «A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century», *Simians, Cyborgs and Women: The Reinvention of Nature,* Nueva York, Routledge, 1991, 149-181.
—, «Anthropocene, Capitalocene, Plantationocene, Chthulucene: Making Kin», *Environmental Humanities* 6, 2015, 159-165.

HERWEIJER, Celine, «8 Ways AI Can Help Save the Planet», *World Economic Forum,* 24 de enero de 2018. Disponible en <https://www.weforum.org/agenda/2018/01/8-ways-ai-can-help-save-the-planet/>.
HOUSE OF COMMONS, «Algorithms in Decision-Making», cuarto informe de la sesión 2017-19, HC351, 23 de mayo de 2018.
ICDPPC (International Conference of Data Protection and Privacy Commissioners), «Declaration on Ethics and Data Protection in Artificial Intelligence», 2018. Disponible en <https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf>.
IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, «Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems», Version 2, IEEE, 2017. Disponible en <http://standards.Ieee.org/develop/indconn/ec/autonomous_systems.html>.
IHDE, Don, *Technology and the Lifeworld: From Garden to Earth,* Bloomington, Indiana University Press, 1990.
JANSEN, Philip, BROADHEAD, Stearns; RODRIGUES, Rowena; WRIGHT, David; BREY, Philip; FOX, Alice y WANG, Ning, «State-of-the-Art Review», borrador del D4.1 presentado a la Comisión Europea el 13 de abril de 2018; SIENNA, un programa de investigación e innovación EU H2020 bajo el acuerdo de subvención núm. 741716.
JOHNSON, Deborah G., «Computer Systems: Moral Entities but not Moral Agents», *Ethics and Information Technology* 8, núm. 4, 2006, 195-204.
KANT, Immanuel, *Lecciones de ética,* Barcelona, Austral, 2013.
KELLEHER, John D. y TIERNEY, Brendan, *Data Science,* Cambridge, MA, MIT Press, 2018.
KHARPAL, Arjun, «Stephen Hawking Says A.I. Could Be 'Worst Event in the History of Our Civilization'», CNBC, 6 de noviembre de 2017. Disponible en <https://www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html>.
KUBRICK, Stanley, dir., *2001: A Space Odyssey,* Beverly Hills, CA, Metro-Goldwyn-Mayer, 1968.
KURZWEIL, Ray, *The Singularity Is Near,* Nueva York, Viking, 2005.
LETA JONES, Meg, «Silencing Bad Bots: Global, Legal and Political Questions for Mean Machine Communication», *Communication Law and Policy* 23, núm. 2, 2018, 159-195.
LIN, Patrick; ABNEY, Keith y BEKEY, George, «Robot Ethics: Mapping the Issues for a Mechanized World», *Artificial Intelligence* 175, 2011, 942-949.
MACINTYRE, Lee C., *Posverdad,* Madrid, Cátedra, 2018.
MARCUSE, Herbert, *El hombre unidimensional,* Barcelona, Ariel, 1964.
MARR, Bernard, «27 Incredible Examples of AI and Machine Learning in Practice», *Forbes,* 30 de abril, 2018. Disponible en <https://www.forbes.com/sites/bernardmarr/2018/04/30/27-incredible-examples-of-ai-and-machine-learn ing-in-practice/#6b37edf27502>.
MCAFEE, Andrew y BRYNJOLFSSON, Erik, *Machine, Platform, Crowd: Harnessing Our Digital Future,* Nueva York, W. W. Norton, 2017.
MILLER, Tim, «Explanation in Artificial Intelligence: Insights from the Social Sciences», *arXiv,* 15 de agosto de 2018. Disponible en <https://arxiv.org/pdf/1706.07269.pdf>.
MOUFFE, Chantal, *Agonística. Pensar el mundo políticamente,* Buenos Aires, Fondo de Cultura Económica, 2014.
NEMITZ, Paul Friedrich, «Constitutional Democracy and Technology in the Age of Artificial Intelligence», Philosophical Transactions of the Royal Society A 376, núm. 2133, 2018. <https://doi.org/10.1098/rsta.2018.0089>.
NOBLE, David F., *La religión de la tecnología,* Barcelona, Paidós, 1999.
PARLAMENTO EUROPEO Y CONSEJO DE LA UNIÓN EUROPEA, «General Data Protection Regulation (GDPR)», 2016. Disponible en <https://eur-lex.europa.eu/legal-content/EN/TXT/?

uri=celex%3A32016R0679>.
REIJERS, Wessel; WRIGHT, David; BREY, Philip; WEBER, Karsten; RODRIGUES, Rowena; O' SULLIVAN, Declan y GORDIJN, Bert, «Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendation», *Science and Engineering Ethics* 24, núm. 5, 2018, 1437-1481.
Royal Society, the, «Portrayals and Perceptions of AI and Why They Matter», 11 de diciembre de 2018. Disponible en <https://royalsociety.org/topics-policy/projects/ai-narratives/>.
RUSHKOFF, Douglas, «Survival of the Richest», *Medium,* 5 de julio, 2018. Disponible en <https://medium.com/s/futurehuman/survival-of-the-richest-9ef6cddd0cc1>.
SAMEK, Wojciech; WIEGAND, Thomas y MÜLLER, Klaus-Robert, «Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models», 2017. Disponible en <https://arxiv.org/pdf/1708.08296.pdf>.
SCHWAB, Katharine, «The Exploitation, Injustice, and Waste Powering Our AI», Fast Company, 18 de septiembre de 2018. Disponible en <https://www.fastcompany.com/90237802/the-exploitation-injustice-and-waste-powering-our-ai>.
SESERI, Rudina, «The Problem with 'Explainable AI'», *Tech Crunch,* 14 de junio de 2018. Disponible en <https://techcrunch.com/2018/06/14/the-problem-with-explainable-ai/?guccounter=1>.
SEARLE, John. R., «Minds, Brains, and Programs», *Behavioral and Brain Sciences* 3, núm. 3, 1980, 417-457.
SHANAHAN, Murray, *The Technological Singularity,* Cambridge, MA, The MIT Press, 2015.
SIAU, Keng y WANG, Weiyu, «Building Trust in Artificial Intelligence, Machine Learning, and Robotics», *Cutter Business Technology Journal* 32, núm. 2, 2018, 4-53.
STOICA, Ion, «A Berkeley View of Systems Challenges for AI». Technical Report No. UCB/EECS-2017- 159, 2017. <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017>.
SULLINS, John, «When Is a Robot a Moral Agent?», *International Review of Information Ethics* 6, 2006, 23-30.
SURUR, «Microsoft Aims to Lie to Their AI to Reduce Sexist Bias», 25 de agosto de 2017. Disponible en <https://mspoweruser.com/microsoft-aims-lie-ai-reduce-sexist-bias/>.
SUZUKI, Yutaka; GALLI, Lisa; IKEDA, Ayaka; ITAKURA, Shoji y KITAZAKI, Michiteru, «Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography», *Scientific Reports* 5, artículo número 15924, 2015. Disponible en <https://www.nature.com/articles/srep15924>.
TEGMARK, Max, *Life 3.0: Being Human in the Age of Artificial Intelligence,* Allen Lane/Penguin Books, 2017.
TURKLE, Sherry, *Alone Together: Why We Expect More from Technology and Less from Each Other,* Nueva York, Basic Books, 2011.
TURNER, Jacob, *Robot Rules: Regulating Artificial Intelligence,* Cham, Palgrave Macmillan, 2019.
UNIVERSITÉ DE MONTRÉAL, «Montréal Declaration Responsible AI», 2017. Disponible en <https://www.montrealdeclaration-responsibleai.com/the-declaration>.
VALLOR, Shannon, *Technology and the Virtues,* Nueva York, Oxford University Press, 2016.
VIGEN, Tyler, *Spurious Correlations,* Nueva York, Hachette Books, 2015.
VILLANI, Cédric, *For a Meaningful Artificial Intelligence: Towards a French and European Strategy,* misión parlamentaria comprendida entre el 8 de septiembre de 2017 y el 8 de marzo de 2018, asignada por el Primer Ministro de Francia, Èdouard Philippe, 2018.
VON SCHOMBERG, René, «Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields», informe de los Servicios de la Comisión Europea, Luxemburgo, Oficina de publicaciones de la Unión Europea, 2011.

VU, Mai- Anh T.; ADALI, Tülay; BA, Demba; BUZSÁKI, György; CARLSON, David; HELLER, Katherine, et al., «A Shared Vision for Machine Learning in Neuroscience», *Journal of Neuroscience* 38, núm. 7, 1601-607, 2018.

WACHTER, Sandra, Brent MITTELSTADT, y Luciano FLORIDI, «Why a Right to Explanation of Automated Decision-Making, Does Not Exist in the General Data Protection Regulation», *International Data Privacy Law,* 2017. Disponible en <http://dx.doi.org/10.2139/ssrn.2903469>.

WALLACH, Wendell y ALLEN, Colin, *Moral Machines: Teaching Robots Right from Wrong,* Oxford, Oxford University Press, 2009.

WELD, Daniel S. y BANSAL, Gagan, «The Challenge of Crafting Intelligible Intelligence», 2018. Disponible en <https://arxiv.org/pdf/1803.04263.pdf>.

WINFIELD, Alan F. T. y JIROTKA, Marina, «The Case for an Ethical Black Box», *Towards Autonomous Robotic Systems,* editado por Yang Gao, Saber Fallah, Yaochu Jin, y Constantina Lekakou (procesos de TAROS 2017, Guildford, UK, July 2017), 2017, 262– 273.

WINIKOFF, Michael, «Towards Trusting Autonomous Systems», *Engineering Multi-Agent Systems,* editado por Amal El Fallah Seghrouchni, Alessandro Ricci y Son Trao, 2018, 3-20.

YAMPOLSKIY, Roman V., «Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach», *Philosophy and Theory of Artificial Inteligence,* editado por Vincent C. Müller, 2013, 289-296.

YEUNG, Karen, «A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework» (un estudio encargado por el Comité del Consejo de Europa a expertos en derechos humanos, y procesamiento de datos automatizados y diferentes formas de inteligencia artificial), MSI- AUT (2018)05, 2018.

ZIMMERMAN, Jess, «What If the Mega- Rich Just Want Rocket Ships to Escape the Earth They Destroy?», *Guardian,* 16 de septiembre de 2015. Disponible en <https://www.theguardian.com/commentisfree/2015/sep/16/mega-rich-rocket-ships-escape-earth>.

ZOU, James, y SCHIEBINGER, Londa, «Design AI So That It's Fair», *Nature* 559, 2018, 324-326.

Otras lecturas recomendadas

ALPAYDIN, Ethem, *Machine Learning,* Cambridge, MA, MIT Press, 2016.

ARENDT, Hannah, *La condición humana,* Barcelona, Paidós, 1993.

ARISTÓTELES, *Aristóteles. Ética nicomáquea. Ética eudemia,* Madrid, Gredos, 1985.

BODDINGTON, Paula, *Towards a Code of Ethics for Artificial Intelligence,* Cham, Springer, 2017.

BODEN, Margaret A., *AI: Its Nature and Future,* Oxford, Oxford University Press, 2016.

BOSTROM, Nick, *Superinteligencia,* Zaragoza, TEELL, 2016.

BRYNJOLFSSON, Erik, y MCAFEE, Andrew, *La segunda era de las máquinas,* Buenos Aires, Temas, 2013.

COECKELBERGH, Mark, *Growing Moral Relations: Critique of Moral Status Ascription,* Nueva York, Palgrave Macmillan, 2012.

CRUTZEN, Paul J., «The 'Anthropocene'», *Earth System Science in the Anthropocene,* editado por Eckart Ehlers y Thomas Krafft, 13-18, Cham, Springer, 2006.

DIGNUM, Virginia, BALDONI, Matteo; BAROGLIO, Cristina; CAON, Maruiyio; CHATILA, Raja; DENNIS, Louise; GÉNOVA, Gonzalo, et al., «Ethics by Design: Necessity or Curse?», Association for the Advancement of Artificial Intelligence (Asociación para el avance de la inteligencia artificial), 2018. Disponible en <http://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_68.pdf>.

DREYFUS, Hubert L., *What Computers Can't Do,* Nueva York, Harper & Row, 1972.

EUROPEAN COMMISION, AI HLEG (High- Level Expert Group on Artificial Intelligence), «Ethics Guidelines for Trustworthy AI», 8 de abril de 2019, Bruselas, Comisión Europea. Disponible en <https://ec.europa.eu/futuriumen/ai-alliance-consultation/guidelines#Top>.

FLORIDI, Luciano; COWLS, Josh; BELTRAMETTI, Monica; CHATILA, Raja; CHAZERAND, Patrice; DIGNUM, Virginia; LUETGE, Christoph; MADELIN, Robert; PAGALLO, Ugo; ROSSI, Francesca; SCHAFER, Burkhard; VALCKE, Peggy y VAYENA, Effy, «AI4People— An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations», *Minds and Machines* 28, núm. 4: 2018, 689-707.

FRANKISH, Keith y RAMSEY, William M., eds. *The Cambridge Handbook of Artificial Intelligence,* Cambridge, Cambridge University Press, 2014.

FRY, Hannah, *Hola mundo: cómo seguir siendo humanos en la era de los algoritmos,* Barcelona, Blackie Books, 2019.

FUCHS, Christian, *Digital Labour and Karl Marx,* Nueva York, Routledge, 2014.

GUNKEL, David, *The Machine Question,* Cambridge, MA, MIT Press, 2012.

HARARI, Yuval Noah, *Homo deus: breve historia del mañana,* Madrid, Debate, 2016.

HARAWAY, Donna, «A Cyborg Manifesto: Science, Technology, and Socialist- Feminism in the Late Twentieth Century», *Simians, Cyborgs and Women: The Reinvention of Nature,* Nueva York, Routledge, 1991, 149-181.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, «Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems», Version

2, IEEE, 2017. Disponible en <http://standards.Ieee.org/develop/indconn/ec/autonomous_systems.html>.

KELLEHER, John D. y TIERNEY, Brendan, *Data Science,* Cambridge, MA, MIT Press, 2018.

NEMITZ, Paul Friedrich, «Constitutional Democracy and Technology in the Age of Artificial Intelligence», *Philosophical Transactions of the Royal Society* A 376, núm. 2133, 2018. Disponible en <https://doi.org/10.1098/rsta. 2018.0089>.

NOBLE, David F., *La religión de la tecnología,* Barcelona, Paidós, 1999.

REIJERS, Wessel; WRIGHT, David; BREY, Philip; WEBER, Karsten; RODRIGUES, Rowena; O'SULLIVAN, Declan y GORDIJN, Bert, «Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendation», *Science and Engineering Ethics* 24, núm. 5, 2018, 1437-1481.

SHELLEY, Mary, *Frankenstein,* Madrid, Anaya, 2010.

TURKLE, Sherry, *Alone Together: Why We Expect More from Technology and Less from Each Other,* Nueva York, Basic Books, 2011.

WALLACH, Wendell y ALLEN, Colin, *Moral Machines: Teaching Robots Right from Wrong,* Oxford, Oxford University Press, 2009.

Título original de la obra:
AI Ethics

Edición en formato digital: 2021

Calle Juan Ignacio Luca de Tena, 15

28027 Madrid

catedra@catedra.com

ISBN ebook: 978-84-376-4217-8

Conversión a formato digital: REGA www.catedra.com